A Case for Simple Tables

Martin A. KOSCHAT

The task of devising and interpreting tables is an integral part of statistical practice. Yet tables receive little attention as a topic of statistical research and statistical education. This neglect seems to be reflected in the quality of the tables that accompany scientific and nonscientific presentations. This article argues that in statistics tables are important tools for communicating information, and hence should receive more attention in statistical research, education, and practice. I discuss basic tables, graphically enhanced tables and, in the context of OLAP, dynamic tables. Examples supporting the case should be of interest to statistical practitioners and educators alike.

KEY WORDS: Data analysis; OLAP; Statistical graphics.

1. INTRODUCTION

A statistical analysis starts with exploration and ends with presentation. Exploration includes gaining an understanding of the data sources, inspecting the data for data integrity as well as unusual features that can guide the development of a formal analysis, and staging the data for further analysis. Presentation includes communicating the findings of the analysis to its intended audience.

Tables and graphs figure prominently in both instances. George Box's (1988) dictum still holds true: "A first analysis of experimental results should, I believe, invariably be conducted using flexible data analytical techniques—looking at graphs and simple statistics—that so far as possible allow the data to 'speak for themselves'. The unexpected phenomena that such an approach often uncovers can be of the greatest importance in shaping and sometimes redirecting the course of an ongoing investigation." This is indeed the course of action initially taken by most statistical practitioners. In the end, academic as well as nonacademic articles and reports that communicate the findings of statistical analyses usually include tables and graphs.

The benefits of the graphical display of information are widely acknowledged, and statistical graphics has been an area of active and lively research for many years. A substantial bibliography of academic research articles, a dedicated major research journal—namely the *Journal of Graphical and Computational Statistics*—and books (Tukey 1977; Mosteller and Tukey 1977; Chambers, Cleveland, Kleiner, and Tukey 1983; Cleveland 1994) that by now are seminal texts in the field of statistics, attest to a robust history and the continued vitality of statistical graphics as an area of research.

In the statistical community, the benefits of tables as a format for displaying information are acknowledged less enthusiastically. It would seem that the value of tables as a format for displaying information is recognized mostly indirectly through the frequency of use. Tables are indeed widely used, and in many statistical reports and research papers more space is devoted to tables than graphs. Yet this obvious prominence is barely reflected in the broader statistical discourse. Tables as displays of information are rarely a topic in statistical education, are rarely a point of discussion in statistical practice and, within the field of statistics, do not appear to receive much, if any, attention from researchers.

An early and noteworthy contribution on the design of statistical tables to the statistical literature is Walker and Durost's (1936) text, considered by many to be a minor masterpiece. Since then it would appear that scholarly articles on this topic have been published mostly by a small handful of researchers, notably Ehrenberg (1981, 1986), Wainer (1992, 1993, 1997a, 1998), and Tufte (2003). The latter two authors also acknowledge tables as a valuable format for communicating information in their books (Tufte 1983, 1990, 1997; Wainer 1997b), but in the statistical community these texts are mostly referred to for their consideration of statistical graphics. Outside statistics, books on document design occasionally have sections on the construction of tables (Bigwood, Spore, and Seely 2003; Harris 2000; Shriver 1997).

The purpose of this article is to address the status quo and contribute to a broader discussion on the value of tables and the attention that tables should receive in education, practice, and research. I perceive there to be a need, as well as an opportunity, and will elaborate on both aspects in the next few sections. This is obviously an ambitious objective fraught with some uncertainty as to its outcome. A a more modest objective is the hope that statistical practitioners and educators will find useful ideas in the examples presented.

2. THE CASE

The prescription for the design of a table is straightforward. Arrange numbers—and it is usually numbers—in parallel *rows* and perpendicular *columns*. A table is a *simple structure* for arranging *numbers*. These two defining attributes—the well-known structure and the use of numbers—also provide the major rationale for using tables.

As to the structure of tables, it is probably moot to speculate whether tables' simple ingenuity is responsible for their wide use or whether it is tables' wide availability and use that have created familiarity and wide acceptance. The fact remains that even users untrained in any quantitative discipline readily comprehend that there is an implied commonality to all the numbers in the same row, often pertaining to a "case," and all the numbers in the same column, often with the meaning of a "variable."

Martin Koschat is Executive Vice President, Information Management, Time/Warner Retail Sales and Marketing, Sports Illustrated Building, Seventh Floor, 135 West 50th Street, New York, NY 10020-1201 (E-mail: martin_koschat@timeinc.com). It is a pleasure to acknowledge the thoughtful comments and helpful suggestions received from Andreas Buja, Anita Hussey, Naomi Robbins, Raja Velu, the editor, associate editor, and two anonymous reviewers.

 DOI: 10.1198/000313005X21429

Magazine	2000 Circ	2000 Ad	2000 Total	2001 Circ	2001 Ad	2001 Total	Circ YOY	Ad YOY	Total YOY
People	442433	723733	1166166	463722	656222	1119944	4.80%	-9.30%	-4.00%
TV Guide	619791	450406	1070198	566793	388880	955673	-8.60%	-13.70%	-10.70%
Time	290988	661094	952082	285400	557623	843023	-1.90%	-15.70%	-11.50%
Sports Illustrated	284636	651179	935815	270750	570981	841731	-4.90%	-12.30%	-10.10%
Better Homes & Gardens	150439	471112	621550	150517	491271	641788	0.10%	4.30%	3.30%
Reader's Digest	314988	274996	589984	314678	265804	580482	-0.10%	-3.30%	-1.60%
Parade	-	582188	582188	-	570482	570482	-	-2.00%	-2.00%
Newsweek	149493	439552	589044	178668	338675	517343	19.50%	-22.90%	-12.20%
Businessweek	57755	572092	629847	60488	394303	454791	4.70%	-31.10%	-27.80%
Good Housekeeping	126196	288484	414681	93931	318333	412264	-25.60%	10.30%	-0.60%
Fortune	57356	476849	534204	55740	333808	389548	-2.80%	-30.00%	-27.10%
Cosmopolitan	86019	260493	346511	105778	273762	379540	23.00%	5.10%	9.50%
Woman's Day	96978	287794	384772	95118	273694	368812	-1.90%	-4.90%	-4.10%
Forbes	56338	434344	490683	57389	305824	363213	1.90%	-29.60%	-26.00%
Family Circle	100259	237782	338041	112716	245715	358431	12.40%	3.30%	6.00%
USA Weekend	-	307775	307775	-	316763	316763	-	2.90%	2.90%
InStyle	59267	231821	291088	62001	238438	300439	4.60%	2.90%	3.20%
Entertainment Weekly	89334	213293	302626	96100	198507	294607	7.60%	-6.90%	-2.60%
Martha Stewart Living	72409	197797	270206	80815	210111	290926	11.60%	6.20%	7.70%
U.S. News & World Rep.	99415	224891	324306	104624	178975	283599	5.20%	-20.40%	-12.60%

Figure 1. Estimated revenue (in $U.S. ,000) in total (Total) and by major source, advertising (Ad) and circulation (Circ), for the 20 leading U.S. consumer magazines for 2000 and 2001 as well as year-over-year (YOY) changes. Source: Advertising Age (www.adage.com). Generating Tool: Microsoft Excel.

This universal understanding of a table's structure, shared by few other statistical constructs, is a simple but powerful argument for using tables.

The act of displaying numbers often gets a bad rap. On the one hand, there is the notion that the sensibilities of the quantitatively challenged need to be protected; this is often a concern in business research. On the other hand, analysts believe that an analysis that starts with basic numbers necessarily needs to yield something more profound than numbers; this is a notion that is often encountered in the sciences. There are, however, good arguments for using numbers in displays, and we mention three.

A numerical display often presents data in their original form and subjects them to minimal intervention from an analyst pursuing a directed investigation. An analysis of data is, of course, nothing but a directed investigation, and a thorough analysis may well require the transformation of original data through graphical constructs or the application of a sophisticated mathematical model. Although such transformations have significant and well-known benefits, the appreciation for these benefits needs to be balanced by a consideration for the bias that such transformations may introduce and that is usually transparent to the end-user of an analysis. The outcome of an analysis is as much driven by the analytical assumptions and choices as by the data. At a minimum, there is a distinct benefit to complementing the presentation of a formal analysis by an informative presentation of the underlying data in their original form or a meaningful and unambiguous numerical summary of such data.

Data presented in numerical form can also be easily manipulated and transformed. The reader can easily take the data as input for a graph or a formal model of her choosing. Numerical displays support and encourage such a flow. It is easy to translate numbers into graphs or parameters of a model. It is much harder, if not impossible, to take the reverse route.

Finally, numbers often—not always—require less of an explanation than glyphs or modeled constructs. In many areas of statistical application, such as business, analysts and analytical clients can easily put the number presented to them into a context where the number has some immediate meaning. In business, it helps, of course, that the raw numbers encountered usually fall into a few standard and well-understood categories. Further, business people often "live" their numbers, particularly if these happen to be sales numbers. Not only do these people fully comprehend the meaning of a sales number, *they want to see it!*

I do not wish to suggest that the presentation of numerical information is uniformly better than a graphical display or the presentation of a formal analysis. I do want to emphasize, however, that there are often good reasons for displaying numerical information in a simply structured format even if only to complement graphs or to support the communication of model-based results.

3. SOME CONSIDERATIONS FOR CONSTRUCTING TABLES

In many computer programs for statistical analysis, raw and derived data tend to be organized in rows and columns, thus naturally forming tables. Analysts often succumb to the temptation to grab such tables in their original form and to drop them into a document or a presentation without further concern for making appropriate adjustments. The results of this approach are often painful to behold as Figure 1, a table taken from an actual pre-

Row Presentation

364513897351.12 364513987351.12 36451389735.112

Column Presentation

364513897351.12
364513987351.12
36451389735.112

Figure 2. The same three numbers are arranged as a row and as a column. The column presentation is better suited to rank-order the three numbers.

sentation to the senior management of a major U.S. publishing house, shows.

This table attempts to compare changes in the revenue breakdown of the 20 leading consumer magazines in the U.S. for two consecutive years. The table fails on several basic dimensions. For the table to be an effective display of data, its entries must be easy to compare. Also, entries and comparisons that are of special interest should be prominently displayed and easy to find. Neither is the case in Figure 1. This simple observation leads to a few basic principles useful in the construction of informative tables. These principles pertain to the choice of rows and columns, the arrangement of rows and columns, the presentation of numbers, and the use of simple graphical elements for structuring the display.

An application of these principles resulted in the table in Figure 3. This table now effectively shows the gestalt of the market comprised of the 20 magazines. It quickly communicates the relative size of each magazine and the relative contribution of each source to total revenues. Most importantly, it clearly communicates that 2001 was a difficult year for magazine publishing, with the difficulty mostly due to a softening of the advertising market.

We use Figure 3 to briefly discuss each of the guiding principles.

3.1 Choice of Columns and Rows

A first step in the construction of a table requires a decision on which entries to arrange in rows and which entries to arrange in columns. In general, numerical comparisons are easier made within columns than within rows. Figure 2 illustrates this point. The same set of three numbers is arranged as a row and also as a column showing that numbers are easier to compare and to rank order in the column presentation.

Apart from supporting the case, the example also indicates why this should be so. The ranking of two numbers is entirely determined by the left-most digit in which the two numbers differ. The assessment of which is larger is then reduced to the inspection of a single digit. The determination of the position of this distinguishing digit and the comparison is easily done in the column presentation even for fairly large numbers, provided, of course, all digits have been properly aligned.

In the consumer magazine example, one may surmise that the comparison of different magazines is probably of greater interest than the comparison of revenue sources. Hence I left the column/row breakdown as it was in the original table.

3.2 Arrangement of Rows and Columns

Of equal importance, albeit for different reasons, are the relative arrangement of rows and the relative arrangement of columns. Rows whose entries one wishes to compare should ideally be displayed close together, and the same holds true for the arrangement of columns. Often the data themselves suggest such a grouping. For example, it may be useful and informative to arrange rows such that the entries in the principal column of interest appear sorted. Also, because we tend to read from left to right and top to bottom, a table's left upper quadrant is likely to receive most of a reader's initial attention. Hence one should consider arranging rows and columns such that the entries of greatest interest fall into the left upper quadrant.

In the redesign of the table in Figure 1 I left the original order of the rows intact because it was determined by each magazine's total revenue in 2001, resulting in a display with the largest and presumably most important magazines in the leading rows. I rearranged the columns so that now all columns pertaining to a particular revenue stream are grouped together, with the columns related to total revenue, arguably the most important revenue figure, moved farthest to the left.

3.3 Presentation of Numbers

It is usually not possible to arrange all entries that one should, or might want to, compare in adjacent and vertically aligned positions. In such instances the reader has to commit, however briefly, at least one of the numbers to be compared to memory. The number of distinct digits that most people retain easily after a single pass is more or less limited to seven (Miller 1956). It is therefore good practice to transform the data such that five digits or fewer represent each table entry, if possible. Often this can be accomplished by adjusting the scale and, by rounding, limiting the number of digits retained. Thus, rounding is an important step with an additional benefit. Usually the left-most digits of a number are more important than the digits to the right. Retaining too many digits hinders the reader from paying attention to the more important digits.

Of course, readers perform some mental rounding on their own, thus focusing on the relevant digits and retaining a compact mental image of numbers. This task is made easier by the addition of commas for every three digits displayed. Generally, it is the responsibility of the table's designer to make this processing as easy as possible by displaying numbers concisely. This means creating a numerical display that retains as few digits as possible but as many as necessary and that adds useful structure such as commas to the digits retained.

In the example here, I expressed all revenue figures in $ mill.—rounded to the nearest million—added commas where appropriate, and took care to assure that corresponding digits were aligned within each column.

3.4 Simple Graphical Elements

A table entry is characterized not only by its numerical value but also by its position within the table. The effectiveness of a table presentation depends in part on how easy it is to determine an entry's positions within the table and to link the entry to its row and column labels. Two simple graphical elements, lining and shading, help as the redesigned table in Figure 3 illustrates.

	Gross Revenue			Advertising Revenue			Circulation Revenue		
	2000	2001	Change	2000	2001	Change	2000	2001	Change
People	1,166	1,120	-4.0 %	724	656	-9.3 %	442	464	4.8 %
TV Guide	1,070	956	-10.7	450	389	-13.7	620	567	-8.6
Time	952	843	-11.5	661	558	-15.7	291	285	-1.9
Sports Illustrated	936	842	-10.1	651	571	-12.3	285	271	-4.9
Better Homes & Gardens	622	642	3.3	471	491	4.3	150	151	0.1
Reader's Digest	590	580	-1.6	275	266	-3.3	315	315	-0.1
Parade	582	570	-2.0	582	570	-2.0	-	-	-
Newsweek	589	517	-12.2	440	339	-22.9	149	179	19.5
Businessweek	630	455	-27.8	572	394	-31.1	58	60	4.7
Good Housekeeping	415	412	-0.6	288	318	10.3	126	94	-25.6
Fortune	534	390	-27.1	477	334	-30.0	57	56	-2.8
Cosmopolitan	347	380	9.5	260	274	5.1	86	106	23.0
Woman's Day	385	369	-4.1	288	274	-4.9	97	95	-1.9
Forbes	491	363	-26.0	434	306	-29.6	56	57	1.9
Family Circle	338	358	6.0	238	246	3.3	100	113	12.4
USA Weekend	308	317	2.9	308	317	2.9	-	-	-
InStyle	291	300	3.2	232	238	2.9	59	62	4.6
Entertainment Weekly	303	295	-2.6	213	199	-6.9	89	96	7.6
Martha Stewart Living	270	291	7.7	198	210	6.2	72	81	11.6
US News & World Report	324	284	-12.6	225	179	-20.4	99	105	5.2

Figure 3. Estimated revenue (in $U.S. 000,000) in total and by source for the 20 leading U.S. consumer magazines for 2000 and 2001 as well as year-over-year changes. Source: Advertising Age (www.adage.com). Generating tool: Microsoft Excel.

Lining refers to the judicious and parsimonious addition of lines to the basic rectangular data display. The emphasis is on "judicious" and "parsimonious" because, as in Figure 1, the problem is often not that there are too few but that there are too many lines. Separating each pair of adjacent rows and columns by a line results in a useless grid that does nothing to help the reader orient herself. On the other hand, the horizontal lines added to the table in Figure 3 define horizontal bands of five rows each, with two complementary benefits. On the one hand, the bands are sufficiently wide and distinct to be easily traceable as the reader's glance moves from left to right. On the other hand, each band is sufficiently narrow to let each row's position be easily determined within the band. If the sole purpose of lining is to help the reader orient herself, the number of rows within each band should be between three and five.

Lines can also be used to group rows with common themes. Assuming the number of rows within each group is sufficiently small, the actual number of rows within each band will then be determined by the subject matter context that defines the groupings, and it may well vary by band.

Shading refers to changing the background color for selected rows and columns. The table in Figure 3 has shaded columns containing revenue changes. In this example, the benefits of shading are two-fold. First, similar to lining, the shading of selected columns creates bands that help determine the column positions of individual entries. Second, shading creates groups of rows or columns that the reader might wish to compare. This is particularly useful in instances where the rows or columns one wishes to compare are not adjacent to one another.

Shading is a frequently used typographical option (Wheildon 1990). It is a delicate addition to a table, and it has to be used cautiously and judiciously. In particular, it is well known that the readability of text deteriorates as the background tint gets too dark. In general, lightly hued background colors are preferable to gray. If color is not an option, careful attention needs to be paid to choosing a gray scale that is dark enough to serve the purpose of structuring the table and that is also light enough so that numbers can be read easily.

There are other typographical elements one can consider with benefits perhaps similar to or complementary to lining and shading. Adherents of the "minimal use of ink" school of thought might argue that in Figure 3 an effect similar to lining could have been achieved simply by increasing the line spacing every five rows. Arguably, an effect similar to shading could have been achieved by, instead of shading selected cells, choosing a font distinct in type, style, or size for the entries in these cells. Figure 6 includes an example of such alternatives.

I would like to emphasize that the implementation of these suggestions are well within the reach of most statistical analysts. Although one could argue that Microsoft Excel is singlehandedly responsible for creating most of the abominations looking like the table in Figure 1, one also needs to acknowledge that this spreadsheet program has the flexibility to help an analyst prepare tables like the one in Figure 3 with a modest amount of effort.

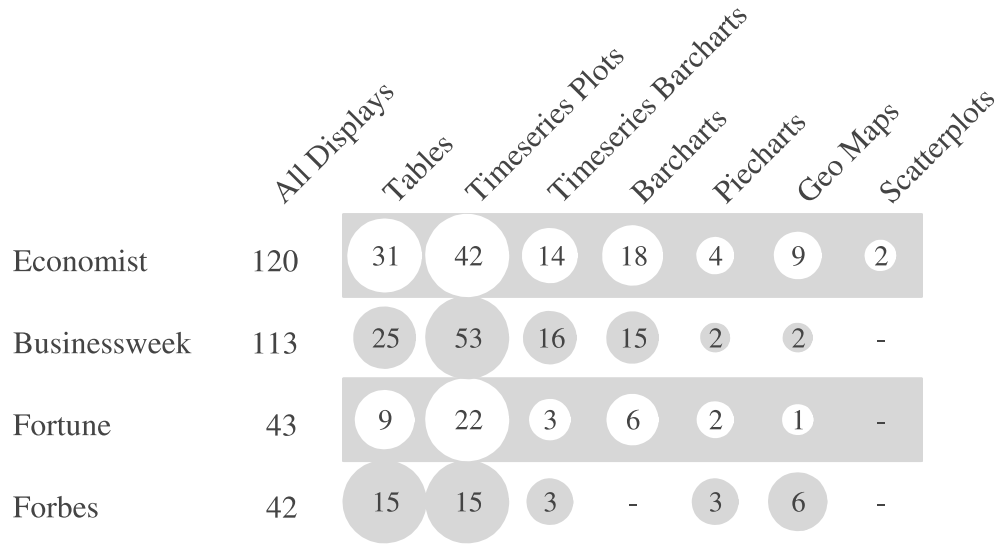

Figure 4. Number of data displays in three arbitrarily chosen issues published in 2003 by major business publications sold in the U.S. Within each row—but not across rows!—circles are proportional in area to frequency counts. Generating tool: S-Plus 6 (2001).

The suggestions of this section are not intended as a comprehensive style guide for the design of tables. Some of the references mentioned in the introduction present interesting and, in specific applications, useful design ideas that have not been covered here. For example, Tufte (2003) shows a table on cancer survival rates that breaks the standard rectangular layout of a simple table to good effect. Also, our suggestions do not provide automatic rules that absolve the analyst from reconciling often-conflicting considerations. However, in my experience these guidelines together with a modest amount of effort significantly improve on the design of the default tables copiously generated by spreadsheet and other data manipulation programs. I hope that these suggestions be used as part of a broader effort that is informed by a variant of the Socratic Principle: *The unexamined table is not worth showing!*

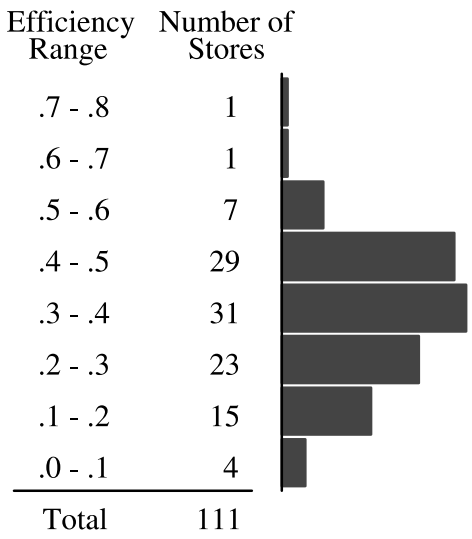

Figure 5. Breakdown of annual sell-through efficiencies (sales/inventory) for a monthly magazine for the 111 stores of a small Midwestern supermarket chain. Generating tool: S-Plus 6 (2001).

4. ADDING GRAPHICAL INFORMATION TO TABLES

A well-crafted table provides a concise presentation of data. If done well, it guides the viewer in the exploration of numerical information, in the process revealing valuable structural insights. I have already noted the value of simple graphical additions for structuring tables. A table's ability to communicate the gestalt of data can be further enhanced by the addition of graphical elements that themselves contain information. Such enhancements may be additions to table entries or additions to the table as a whole. A few examples will illustrate the point. (I do not lay any claim to originality and vaguely—hence no references—recall having seen variants of the graphics used in these examples elsewhere.)

Figure 4 includes a table that records for each of four popular business publications sold in the U.S. the frequency of several data displays used by the magazines. Within each row—but not across rows!—a circle proportional in area to its numerical value surrounds each entry. In conjunction with horizontal shading, the circles induce a comparison of the share of each display within each magazine. The graphical enhancement quickly communicates the relative prominence of each display in each publication. Also, note the angled column labels. Such angled labels can be fairly long, but are still readable without a need for turning the table.

The second example is motivated by common statistical practice. In order to capture the variation in a magazine's sell-through efficiency—calculated as the ratio of sales and inventory—across the stores of a small supermarket chain, we broke up the efficiency range into contiguous intervals of equal length and counted the number of stores falling into each interval. Figure 5 shows the resulting table that has been enhanced by horizontal bars proportional in length to the respective frequency counts. The resulting bar chart is, of course, nothing but the rotated mirror image of a standard histogram.

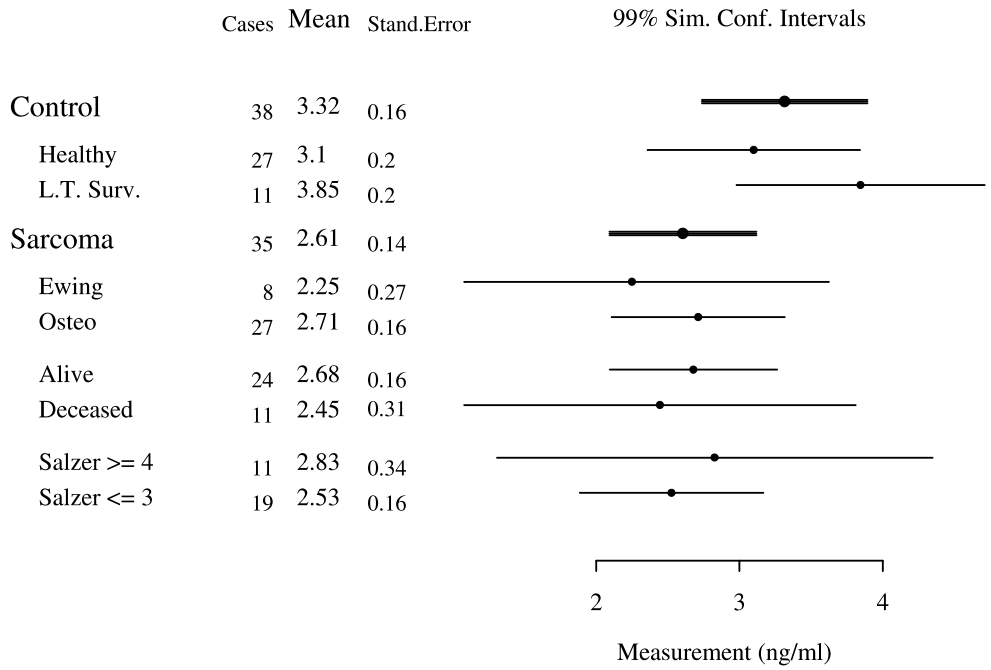

Figure 6. Numerical summaries of $p185^{HER-2}$ concentration for paired partitions of subgroups of patients and control subjects (source: Holzer et al. 2005). Also displayed are simultaneous confidence intervals based on the Bonferroni principle (Bickel and Doksum 2000). Generating tool: S-Plus 6 (2001).

The tabulation made explicit in Figure 5 is an essential step in the construction of a histogram. This frequency tabulation, when first proposed more than 300 years ago by John Graunt (see, e.g., Tapia and Thompson 1978), was duly acknowledged as an important and original scientific accomplishment. In a standard histogram this step is implicit with the result that, in my experience, most readers—even those who have received statistical training in the form of an introductory statistics course—have difficulties interpreting them. On the other hand, few people have difficulties interpreting the table in Figure 5 and *subsequently* interpreting the accompanying bar chart.

The final example shows how tables that present statistical results can be graphically enhanced. Summaries resulting from statistical models are typically presented as tables containing parameter estimates and associated standard errors. The intent of such a presentation is often to invite a comparison of selected parameters to determine whether there is evidence that they are different.

A simple and robust method for the comparison of pairs of parameters is based on an inspection of the corresponding confidence intervals. If these intervals do not overlap it may be taken as evidence that the underlying parameters are indeed different. Hence there is value in adding confidence intervals to the display. Adding the confidence intervals as two additional numeric columns, however, results in an imperfect display. Because a comparison of confidence intervals requires a comparison of the *upper* endpoint of one with the *lower* endpoint of the other interval, a comparison of numbers in different columns would be required. A graphical addition to the base table provides a useful enhancement of the numerical display.

Figure 6 shows a table that includes the number of cases, the means and the standard errors for the blood serum levels of a cancer marker for subgroups of a sample comprised of patients and control individuals. In this table I used different font size and different line spacing to structure the row presentation. In the presentation of the numerical table entries I also broke the horizontal alignment. The vertical offset of the numbers within each row improves item identification by making it easier to determine where one number ends and the next number starts.

The distinguishing feature of this table is the addition of simultaneous confidence intervals based on the Bonferroni principle. The joint coverage probability of these intervals exceeds 99%. As can be seen, these intervals all overlap and therefore do not provide evidence that these means differ from each other. (Denote by $U_k, k = 1, \ldots, K$ confidence intervals whose probability of *jointly* covering corresponding parameters p_k equals or exceeds $1 - \alpha$. Note, that if the individual coverage probability of each interval is at least $1 - \alpha/K$, the Bonferroni principle guarantees that the joint coverage probability equals or exceeds $1 - \alpha$. The decision rule "Do not reject $H_0 : p_1 = \cdots = p_K$ if the U_k contain at least one common point; otherwise reject" defines a level α test.)

Here the table and the graphical addition effectively complement each other. On the one hand, the graphical display helps make the desired comparison. On the other hand, the table readily displays all the information a reader would need should she choose to change the confidence level and with it the confidence intervals.

5. DYNAMIC TABLES

As the volume and the complexity of data increase, a single simple table will usually no longer do the data justice. One could attempt to deal with the problem by devising some appropriately complex multiway layout. Such an approach may, however, defeat its intended purpose by resulting in a data display that lacks

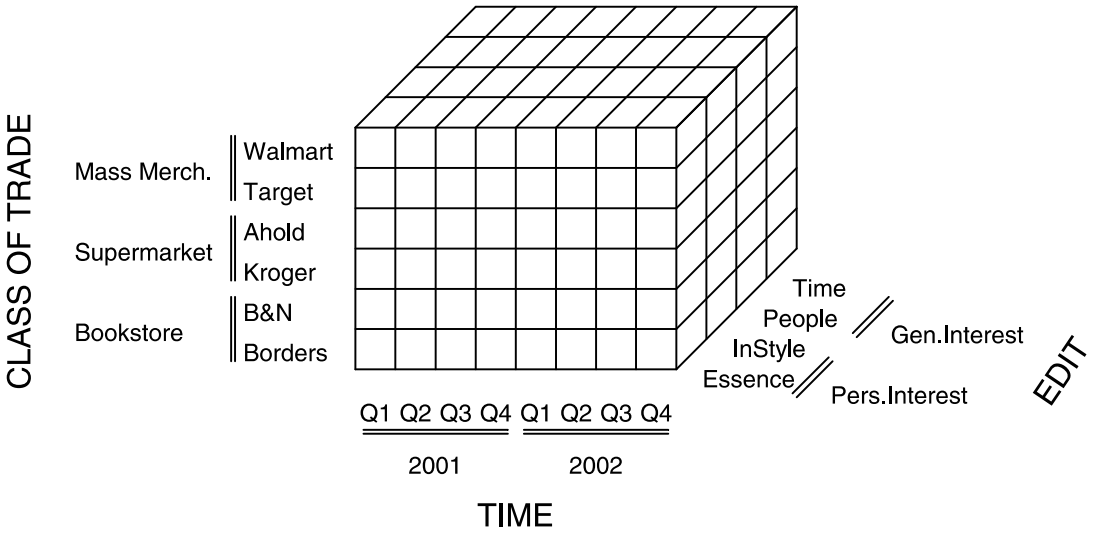

Figure 7. Example of an OLAP cube with three dimensions and two variables for each dimension. The contents of the cells may be measurements such as inventory, sales or revenues.

the self-explanatory immediacy of a simple table. An alternative tactic consists of providing an environment that permits an analyst to create and organize a multitude of simple tables in real time. Through selection and modification, the analyst can then guide herself to a small set of tables that capture and highlight interesting structure in the data.

Such a proposal leads to interesting questions for statistical methodological research. Fortunately the implementation of this proposal does not have to await the outcome of such research because systems that are viable for analytical practice are already available. Within decision-support software, there are programs developed around the concept of OLAP (on-line application programming). From the user's point of view the three essential elements of OLAP are its structuring of data, a user interface that exploits this structure, and the speed with which the data can be accessed and manipulated.

In OLAP, data—usually summary measurements—are organized in a multiway layout along dimensions where each dimension consists of hierarchically structured or nested categorical variables, each with a distinct set of levels (Shoshani 2003). Figure 7 provides an example, again from magazine publishing, that illustrates the concept. The measurement could be inventory, unit sales, revenues, or profits for a particular magazine during a fiscal quarter in a particular retail chain. These measurements are organized along the dimensions "Class of Trade," "Time," and "Editorial Category." In our example, each dimension has two variables. In general, there can be more than two variables per dimension, and there can be more than three dimensions. In OLAP, data organized in this manner are, for obvious reasons, referred to as a "cube."

Through a dedicated interface, the analyst interacts with a cube by selecting subranges and summarizing over subranges to generate subcubes. The analytical objective is to generate informative subcubes of low dimensionality. Preferred low-dimensional subcubes are tables.

In an application like the one described in Figure 7 the objective might be to track changes in magazine sales over a particular time period for a particular class of trade and, if changes occur, to identify the sources of these changes. The analyst often starts with a high level summary view and, by selecting and filtering, "drills down" to interesting and revealing views.

The analyst interacts with the data with the help of a user interface that is designed around the data structure. Figure 8(a) shows an example of a Web-enabled user interface from a commercial OLAP tool (PowerPlay by Cognos, 2003). A table is created using three control buttons for choosing the measure one wishes to display, as well as the horizontal and vertical variables by which the measure is broken down. By clicking on one of these control buttons the user is presented with a menu from which to choose the dimension, as well as the variable of each dimension. In addition, each dimension has a control button that allows filtering by the levels of each associated variable.

Working off a cube with a structure similar to the one shown in Figure 7 and starting with the view shown in Figure 8(a) I decided to investigate what over the course of 2002 happened in bookstores. I filtered on the "Time" dimension to select the year 2002 and the "Class of Trade" dimension to select Bookstores. I chose editorial categories and quarters as the horizontal and vertical variables to produce the sale table shown in Figure 8(b).

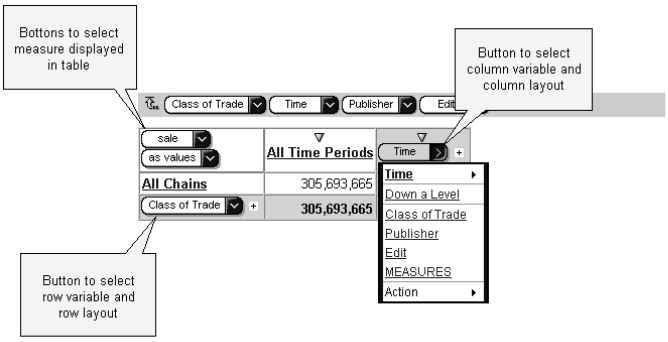

Figure 8a. Commercial OLAP interface (Powerplay by Cognos 2003) to a cube that captures magazine unit sales broken down by the variables associated with the four dimensions "Class of Trade," "Time," "Publisher," and "Edit."

Bookstore | 2002 | Publisher | Edit

sale / as values	2002 Q 1	2002 Q 2	2002 Q 3	2002 Q 4	2002
General Interest	3,863,321	3,343,163	3,968,394	3,763,075	**14,937,953**
Personal Interest	1,681,309	1,440,441	1,688,862	1,538,580	**6,349,192**
Special Interest	3,446,723	2,960,981	3,392,646	3,412,388	**13,212,738**
Uncoded	603,745	511,090	584,040	520,737	**2,219,612**
Edit	**9,595,098**	**8,255,675**	**9,633,942**	**9,234,780**	**36,719,495**

Figure 8b. Unit sales in 2002 in Bookstores broken down by editorial categories and quarters.

It is noteworthy to emphasize the speed with which this activity could be performed; it took less than a minute.

A quick inspection of this table reveals a marked drop in unit sales during the second quarter. At this point one may wish to investigate whether this sales fall-off can be traced to specific sources such as particular editorial categories. This is easily done with some minor adjustments to the sale table. Figure 8(c) shows the sale table with the rows ranked by annual sales volume and the raw sales numbers replaced by each column's row percentage, which is nothing other than each editorial category's market share. Because these shares did not change much across these four quarters, we can conclude that all editorial categories experienced comparable sales losses during the second quarter.

The analysts may now wish to investigate whether the observed sales drop was a singular event or a reflection of inherent seasonality. Displaying quarterly data for more than one year helps address this question. Alternatively, the analyst might want to investigate whether the sales fall-off can be traced to specific stores. If the underlying cube contains store-level detail she can easily drill down to perform the required store level analysis. Again, with a suitable OLAP interface such as the one shown in these examples the necessary comparisons can be easily performed in real-time.

The OLAP interface example shown here is just one commercially available example. Several of the established statistical software vendors now offer OLAP add-on functionality to their statistical offerings. Also, the ubiquitous Microsoft Excel with its Pivot Table provides OLAP functionality.

The criteria for designing a good OLAP interface are similar to the ones laid out by Koschat and Swayne (1996) in a slightly different context. The interface should include a visual representation of every dimension and every variable so that variables and levels are easily available for selection, and so that feedback is provided to the analyst about the current presentation of data. On the other hand, the interface should not be cluttered with irrelevant information. Additionally, if analysts go down analytical dead ends, it should always be easy for them to retrace their steps (Norman 1988). Overall, the interface design should strive for the principles of "direct manipulation" (Schneiderman 1998), which give the analyst a sensation of interacting directly with the data rather than with a computer or a system.

Bookstore | 2002 | Publisher | Edit

sale / as % of column total	2002 Q 1	2002 Q 2	2002 Q 3	2002 Q 4	2002
General Interest	40.26%	40.50%	41.19%	40.75%	**40.68%**
Special Interest	35.92%	35.87%	35.22%	36.95%	**35.98%**
Personal Interest	17.52%	17.45%	17.53%	16.66%	**17.29%**
Uncoded	6.29%	6.19%	6.06%	5.64%	**6.04%**
Edit	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**

Figure 8c. Market share for each editorial category by quarter.

6. DISCUSSION

At the outset, I do not wish to suggest that the mere display of data is a substitute for a comprehensive statistical analysis, which includes statistical inference. Also, I do not claim that a table is always the preferred display of data. I would like to argue, however, that tables should receive more attention within the statistical discipline. In particular, tables can and should receive more thoughtful consideration in statistical practice, statistical education, and statistical research.

It would appear that not all is well in the practice of displaying information. Figure 4 (p. 35) provides an interesting commentary on the type of data displays that the readership of a broad cross-section of business publications is exposed to. These readers tend to be well educated and should comprise a fairly sophisticated audience. Starting with graphs, their seeming abundance is misleading because most graphs tend to be time series plots and time series bar charts. This is a rather modest subset of the options available for a meaningful graphical display of information. About the only positive in Figure 4 is the observation that nowadays the number of pie charts used by all publications is mercifully small.

Even though the count of tables is uniformly smaller than that of graphs, the amount of space devoted to the display of tables is larger than that devoted to the display of graphs. The quality of the table displays varies. Not surprisingly, tables that appear as a regular feature in most issues of a magazine such as stock tables are usually carefully laid out while one-of-a-kind tables are often hard to parse with plenty of room for improvement. Overall, it seems that the everyday practice of data display lags the lofty expectations established by a quarter of a century worth of systematic research on displaying and communicating information.

Information displays devised for what might be expected to be a more knowledgeable audience—namely statisticians—do not necessarily fare better either. Gelman, Pasorica, and Dodhia (2002) point out that the principal means of displaying scientific findings in statistical research journals are often-perfunctory tables, and the authors suggest graphical displays as an alternative.

It is tempting to blame statistics education for this less-than-perfect state of affairs, and it is hard to see how one can resist this temptation. I recently had the opportunity to review 11 popular texts used in introductory business statistics courses for their treatment of the display of information. These texts spent between 1.2% and 8.9% of their pages on this topic. Most texts simply confined themselves to introducing the standard graphical displays. Only two of the texts actually discussed *principles* of good graphical design. These two texts also explicitly acknowledged tables as a means of communicating information

but did not provide any guidance on how to construct informative tables. A third text had a short but serviceable section on Pivot Tables. This state of affairs, in my opinion, calls for improvement.

It is argued that graphs and certainly tables are simple constructs akin to hammers and chisels that do not warrant the same amount of time as the more complicated tools of inference which are, after all, also part of the curriculum. It is certainly true that the principle underlying the design of graphs and tables (like that of hammers and chisels) is quickly communicated. The proper use of these simple tools on the other hand may well require dedicated and sustained training. Significant benefits could be reaped if in introductory statistics courses more time were spent on the proper design and use of graphs and tables. It is too much to hope that such a focus would yield the next John Tukey, Edward Tufte, or Howard Wainer (or Leonardo, for that matter) but it stands to reason that it would improve the quality of data displays that are routinely generated in statistical practice, and that it would beneficially raise expectations regarding a meaningful display of information.

I suspect that, for obvious reasons, academic educators gravitate in their choice of course topics toward those that readily tie into active research. It therefore needs to be emphasized that dynamic tables are the focus of active research. Though OLAP is not the outgrowth of traditional statistical research, it is a subject of active research that readily connects with the statistical research tradition. This connection is the result of OLAP's philosophy of data access, as well as the structure of the underlying data.

Regarding the philosophy of data access and presentation, there is a well-established and successful precedent in statistical graphics of representing complex data by using simple displays that can be easily manipulated. The philosophy driving this research is captured in the visualization programs XGobi (Swayne, Buja, and Cook 1998) and GGobi (Swayne, Temple Lang, Buja, and Cook 2003). In these programs the user generates through a user-friendly and intuitive GUI in real time multiple simple views, such as dot plots, scatterplots and time series plots, that represent different views of complex, high-dimensional data. The information contained in these simple views is enhanced by functions such as linking and painting, which allow the simple displays to jointly represent a more complex whole. In addition, the programs include search utilities based on concepts such as the Grand Tour (Asimov 1985) and Projection Pursuit (Cook, Buja, Cabrera, and Hurley 1995) which help steer the analyst towards views that could be interesting in the specific analytical context.

While these programs are concerned with graphs, an OLAP tool deals with tables. Other than that, there is a remarkable commonality between these utilities. It would seem obvious that many of the thoughts and insights that informed the development of dynamic graphics could be beneficially applied to the development of OLAP. For example, it should be an interesting research proposal to devise search utilities similar to the Grand Tour or Projection Pursuit that help an analyst steer toward interesting and meaningful tables.

OLAP has a second point of contact with statistical research. Linear models (e.g., Rao and Toutenburg 1999) provided a rich platform for inference in the analysis of multiway layouts, the underlying data structure of OLAP. This holds out the prospect for utilities that, bundled with OLAP, may help reconcile an analyst's belief that there is structure in a particular view of the data with formal statistical inference. This reconciliation has traditionally been a challenge for statistical graphical analysis.

Thus OLAP provides not only an interface to data but it has the potential of supporting a broad array of tools for interfacing with information. This ability to interface with information quickly and efficiently and to produce views of the data in a format that is easily understood will be critical in the years to come. In most organizations the amount of data generated grows at a rapid clip while the number of analysts dealing with this data flow tends to stay constant. Analysts and statisticians will need to become more efficient. Concepts such as OLAP will be critical in this effort and it stands to reason that OLAP should therefore be an integral part of a statistics curriculum.

[Received March 2004. Revised October 2004.]

REFERENCES

Asimov, D. (1985), "The Grand Tour: A Tool for Viewing Multidimensional Data," *SIAM Journal of Scientific and Statistical Computing*, 6, 128–143.

Bickel, P. J., and Doksum K. A. (2000), *Mathematical Statistics: Basic Ideas and Selected Topics I*, Upper Saddle River, NJ: Pearson Education.

Bigwood, S., Spore, M., and Seely, J. (2003), *Presenting Numbers, Tables and Charts*, Oxford, UK: Oxford University Press.

Box, G. E. P. (1988), "Signal-to-Noise-Ratios, Performance Criteria, and Transformations," *Technometrics*, 30, 1–17.

Chambers, J., Cleveland, W. S., Kleiner, B., and Tukey, P. (1983), *Graphical Methods for Data Analysis*, Belmont, CA: Wadsworth.

Cleveland W. S. (1994), *The Elements of Graphing Data*, Summit, NJ: Hobart Press.

Cook, D., Buja, A., Cabrera, J., and Hurley, C. (1995), "Grand Tour and Projection Pursuit," *Journal of Computational and Graphical Statistics*, 4, 155–172.

Ehrenberg, A. S. C. (1981), "The Problem of Numeracy," *The American Statistician*, 35, 67–71.

——— (1986), "Reading a Table: An Example," *Applied Statistics*, 35, 237–244.

Gelman, A., Pasarica, C., and Dodhia, R. (2002), "Let's Practice What We Preach: Turning Tables into Graphs," *The American Statistician*, 56, 121–131.

Harris, R. L. (2000), *Information Graphics: A Comprehensive Illustrated Reference*, Oxford, UK: Oxford University Press.

Holzer, G., Pfandlsteiner, T., Koschat, M. A., Blahovec, H., Trieb, K., and Kotz, R. (2005), "Soluble p185$^{\mathrm{Her}-2}$ in Malignant Bone Tumors," *Pediatric Blood and Cancer*, 44, 163–166.

Koschat, M. A., and Swayne, D.F. (1996), "Interactive Graphical Methods in the Analysis of Customer Panel Data" (with discussion), *Journal of Business and Economic Statistics*, 14, 113–132.

Miller, G. A. (1956), "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information," *The Psychological Review*, 62, 81–97.

Mosteller, F., and Tukey J. W. (1977), *Data Analysis and Regression*, Reading, MA: Addison-Wesley.

Norman, D. A. (1988), *The Psychology of Everyday Things*, New York, NY: Basic Books.

Powerplay by Cognos (2003), Palo Alto, CA: Hewlett-Packard Corporation.

Rao, C. R., and Toutenburg, H. (1999), *Linear Models: Least Squares and Alternatives*, New York: Springer.

S-Plus 6.0 (1988–2001), Seattle, WA: Insightful Corporation.

Schneiderman, B. (1998), *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (3rd ed.), Boston, MA: Addison-Wesley.

Shoshani, A. (2003), "Multidimensionality in Statistical, OLAP and Scientific Databases," in *Multidimensional Databases: Problems and Solutions*, ed. M. Rafanelli, Hershey, PA: Idea Group Publishing.

Shriver, K. A. (1997), *Dynamics in Document Design: Creating Texts for Readers*, New York: Wiley.

Swayne, D. F., Cook, D., Buja, A. (1998), "XGobi: Interactive Dynamic Graphics in the X Window System," *Journal of Computational and Graphical Statistics*, 7, 113–130.

Swayne, D. F., Temple Lang, D., Buja, A., and Cook, D. (2003), "GGobi: Evolving from XGobi into an Extensible Framework for Interactive Data Visualization," *Computational Statistics and Data Analysis*, 43, 423–444.

Tapia, R. A., and Thompson, J. R. (1978), *Nonparametric Probability Density Estimation*, Baltimore: The Johns Hopkins University Press.

Tufte, E. R. (1983), *The Visual Display of Quantitative Information*, Cheshire, CT: Graphics Press.

——— (1990), *Envisioning Information*, Cheshire, CT: Graphics Press.

——— (1997), *Visual Explanations*, Cheshire, CT: Graphics Press.

——— (2003), *The Cognitive Style of PowerPoint*, Cheshire, CT: Graphics Press.

Tukey, J. W. (1977), *Exploratory Data Analysis*, Reading, MA: Addison-Wesley.

Walker, H. M., and Durost, W. N. (1936), *Statistical Tables: Their Structure and Use*, New York, NY: Bureau of Publications, Teachers College, Columbia University.

Wainer, H. (1992), "Understanding Graphs and Tables," *Educational Researcher*, 21, 14–23.

——— (1993), "Tabular Presentation," *Chance*, 6, 52–56.

——— (1997a), "Improving Tabular Displays, with NAEP Tables as Examples and Inspirations," *Journal of Educational and Behavioral Statistics*, 22, 1–30.

——— (1997b), *Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot*, New York: Springer.

——— (1998), "Rounding Tables," *Chance*, 11, 46–50.

Wheildon, C. (1990), *Communicating or Just Making Pretty Shapes—A Study of the Validity—or Otherwise—of Some Elements of Typographic Design* (3rd ed.), North Sydney, Australia: Newspaper Advertising Bureau of Australia.