Behavior Research Methods, Instruments, & Computers
1993, 25 (2), 250-256

A picture is worth a thousand *p* values: On the irrelevance of hypothesis testing in the microcomputer age

GEOFFREY R. LOFTUS
University of Washington, Seattle, Washington

Hypothesis testing, while by far the most common statistical technique for generating conclusions from data, is nonetheless not very informative. It emphasizes a banal and confusing question ("Is it true that some set of population means are not all identical to one another?") whose answer is, in a mathematical sense, almost inevitably known ("No"). Hypothesis testing, as it is customarily implemented, ignores two issues that are generally much more interesting, important, and relevant: What is the *pattern* of population means over conditions, and what are the magnitudes of various variability measures (e.g., standard errors of the mean, estimates of population standard deviations)? The simple expedient of plotting relevant sample statistics with associated variability bars is a substantially better way of conveying the results of an experiment. In today's microcomputer environment, there are many relatively cheap and easily available applications that allow one to do this. I make some brief, informal comments about some of these applications.

I want to make two main points in this article. First, hypothesis testing is overrated, overused, and practically useless as a means of illuminating what the data in some experiment are trying to tell us. Second, graphical presentation methods are a much better way to provide such illumination, particularly given the ease with which present computer technology allows such methods to be implemented.

THE ENDURING TYRANNY OF HYPOTHESIS TESTING

In 1962, upon ending his editorship of the august *Journal of Experimental Psychology*, Arthur Melton wrote an editorial in which he summarized criteria used by the journal for accepting manuscripts. These criteria revolved heavily around hypothesis testing. In particular, Melton noted that (1) articles in which the null hypothesis was not rejected were almost never published and (2) rejection at the .05 significance level was rarely adequate for acceptance; rather, rejection at the .01 level was typically required.

Melton's editorial blessed a practice that had already become widespread within the social sciences: the use of hypothesis testing as a necessary (and almost sufficient) technique for data analysis. This practice has not changed much in the intervening 30 years; today, hypothesis testing is the primary means of inferring conclusions from data in over 90% of the articles in the major psychology journals.

Hypothesis testing provides the illusion of scientific objectivity by sanctifying an arbitrary probability ($p = .05$) of incorrectly rejecting some null hypothesis that almost inevitably is known apriori to be false (see Bakan, 1966; Gigerenzer et al., 1989; Loftus, 1991; Nunnally, 1960, for elucidations of this argument).[1] Hypothesis testing, as normally implemented, provides virtually no information about two critical aspects of an experiment: the degree of experimental power and the relationship of a set of population parameters (typically population means)[2] to one another. I will argue that the simple expedient of presenting a figure that depicts sample means, along with relevant error bars (a procedure to which I will refer for expositional convenience as the *plot-plus-error-bar*, or PPE, procedure) provides essentially all the information provided by a hypothesis-testing procedure, plus additional information. Furthermore, the information that is shared by the hypothesis testing and PPE procedures is generally uninteresting and unimportant, whereas the additional information provided by the PPE procedure is generally interesting and important.

There are many reasons why hypothesis testing originally became the default data analysis technique in the social sciences (see discussions by Cohen, 1990; Gigerenzer et al., 1989; Loftus, 1991). One of them is that it has generally been easy to do it. You plug raw data into a computer program and out comes a *z* or a *t* or an *F* value that tells you everything you need for writing your article. (As Cohen, 1990, astutely points out, some peddlers of statistical software packages have gone so far as to hawk their wares by claiming, correctly, that you do not have to even understand statistics in order to use the application.)

The writing of this manuscript was supported by an NIMH grant to G. Loftus. Requests for reprints may be addressed to the author at the Department of Psychology, University of Washington, Department of Psychology, Seattle, WA 98195 (e-mail: gloftus@u.washington.edu).

In contrast to the relative ease of hypothesis testing, making plots with standard errors had, until recent years, been rather tedious. First you had to buy graph paper, pencils, a pencil sharpener, and many erasers. Then you had to spend considerable time just to make a rough preliminary plot. Then you had to take your rough plot to some expensive graphic artist, typically far across campus in the Medical School or somewhere, and wait a week or so for the final result. If you changed your mind about what you wanted to plot, you had to cycle through the whole procedure all over again. There was little in the way of immediate feedback, and the process was not fun.

In the past decade, however, things have changed dramatically. With the explosion of computer graphics, cut-and-paste procedures, and cheap graphing applications, it is very easy to present data as a plot, or a collection of plots, rather than as a compendium of F ratios. That is what we should be doing.

Two Romans à Clef

In this section, I will tell two stories that are meant to illustrate the relationship between the hypothesis testing and PPE procedures. In these stories, the names, experiments, and data have been changed in order to deter hurt feelings, embarrassment, and general professional acrimony.

1. The Time Course of Visual Information Acquisition

A couple of years ago, a cognitive psychologist named Julia Loeb submitted a manuscript to the *Journal of Important Results (JIR)*. Loeb was interested in perceptual encoding of, and memory for, simple dot matrices. Her task was straightforward: on each of many trials, a subject saw a stimulus consisting of four dots embedded in four randomly selected cells of an $n \times n$ matrix. Following the matrix's offset, the subject was required to reproduce the dots' positions.

Loeb's design incorporated three independent variables (all within subjects). First, the stimulus was shown at one of eight exposure durations. There were also two levels of stimulus uncertainty, and two levels of verbal encoding/no verbal encoding (for purposes of today's discussion, a detailed description of these variables is not necessary). Loeb ran 10 subjects in her experiment.

Loeb had developed a theory that implied the following. First, task performance (proportion of correctly located dots) should be exponentially related to exposure duration. Thus, if d is exposure duration, and p is performance, the equation

$$p = (1.0 - e^{-d/c}) \qquad (1)$$

should describe the relation between them (here c is a constant). The second implication of Loeb's theory was that both more uncertain stimuli and lack of verbal encoding should lead to poorer performance.

To examine her results, Loeb planned to plot probability correct, p, as a function of exposure duration, d, and determine the degree to which the resulting curves could be fit by Equation 1. As she was starting to do so, however, she realized that by expressing performance not in terms of raw proportion correct, p, but instead in terms of the transformed score,

$$P = -\ln(1-p),$$

the resulting curves relating performance to duration should be linear rather than exponential. That is, with the use of P rather than p, Equation 1 becomes,

$$P = d/c. \qquad (2)$$

Loeb, a very visually-oriented person, decided that linear functions of the sort described by Equation 2 are easier to assess, comprehend, and compare than are exponential functions of the sort described by Equation 1. Because she could see no drawbacks associated with expressing performance in terms of P rather than p, that is what she did.

Her data, which are reproduced in Figure 1, confirmed her predictions quite nicely. Each panel shows performance as a function of exposure duration. The two curves in each panel represent the two stimulus-uncertainty levels. The two panels show data for the two encoding-strategy levels. For each curve, the data points represent the condition means along with the relevant standard error bar, and the solid line represents the best-fitting linear function. Loeb described a number of other interesting and important aspects of the Figure 1 data having to do with the exact relationships among the slopes of the four functions, but I will skip a discussion of these aspects, for they are not relevant to today's story.

Hypothesis testing as an alternative to Figure 1. The *JIR* reviewers were quite positive about Loeb's manuscript, and the editor accepted it with minor revisions. However, at the very last stage of the editor's interaction with Loeb—as part of the normally benign correspondence in which is enclosed the green to-be-signed document

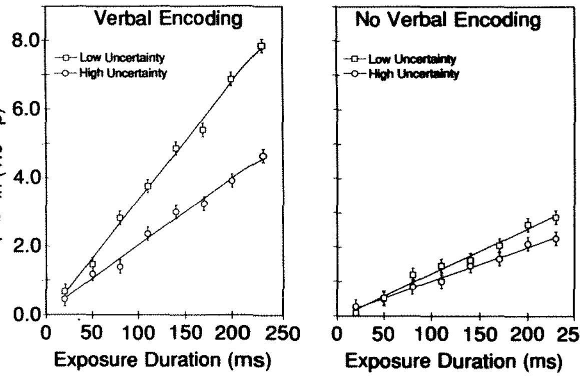

Figure 1. Loeb's data: Mean memory performance plotted as functions of exposure duration. The error bar around each data point is the standard error of the mean. Solid lines represent best-fitting linear functions.

transferring copyright to the journal—a snag occurred. In his letter, the editor added, "In my final reading of your manuscript, I noticed that you didn't do any hypothesis testing on the Figure 1 data. Please include such tests, along with the relevant F values in the final version of your manuscript." Because this was Loeb's tenure year, she didn't want to make any fuss that might endanger her manuscript's publication, so she dutifully added the following paragraph to her results section.

> An 8×2×2 repeated measures ANOVA revealed a main effect of exposure duration, $F(7,63) = 354.49, p < .05$, a main effect of stimulus uncertainty, $F(1,9) = 16.02, p < .05$, and a main effect of encoding strategy, $F(1,9) = 121.33, p < .05$. The interactions of exposure duration with uncertainty and encoding strategy were both significant, $Fs(7,63) = 82.23$ and 77.90, respectively, both $ps < .05$. The uncertainty × strategy interaction was significant, $F(1,9) = 24.98, p < .05$. The duration × uncertainty × strategy interaction was significant, $F(7,63) = 13.23$, $p < .05$.

The *JIR* editor was so pleased with this paragraph that he suggested Figure 1, which he said was now redundant, be removed (thereby following a long tradition of journal editors who, pressured by cost-of-paper-conscious publishers, are always suggesting that figures be removed). This time, however, Loeb stuck to her guns and, in due course, both Figure 1 and the paragraph reproduced above were published.

What's Wrong with this Story? Let's step back a moment and look at the big picture. What is important to know about Loeb's data? Simply by looking at Figure 1, we can infer quite a bit. First, because the predicted linear functions fall within the error bars, we conclude that linearity describes the individual curves quite adequately. Second, because the confidence intervals themselves are quite small, we conclude that the data enjoy substantial statistical power: That is, any deviation of the relevant population means from observed sample means (and thus any departure from linearity on the part of the actual population curves) must be small. Third, by comparing the two curves within each panel, we can conclude that higher uncertainty leads to poorer performance. Finally, by comparing the curves across panels, we can conclude that preventing verbal encoding leads to poorer performance. The last two conclusions are unambiguous, given the large condition differences relative to the small confidence intervals.

Figure 1 also allows some utterly banal conclusions. For instance, we can easily conclude that, within a given curve, the eight population means corresponding to the eight exposure-duration conditions are not identical to one another; if they were, then, given the size of the confidence intervals, the sample means could not plausibly differ from one another by as much as they do. We could make analogous conclusions about the other variables. I characterize such conclusions as banal because we know *a priori* that they must be true. No set of real-valued condition population means can be identical to an infinite number of decimal places. They *must* differ. So why is it interesting to conclude that they do? It is not. What is interesting is not that the population means differ from one another, but rather what the *pattern* of population means is. Here, for instance, it is important to be able to conclude that the relation between population means and exposure durations is linear.

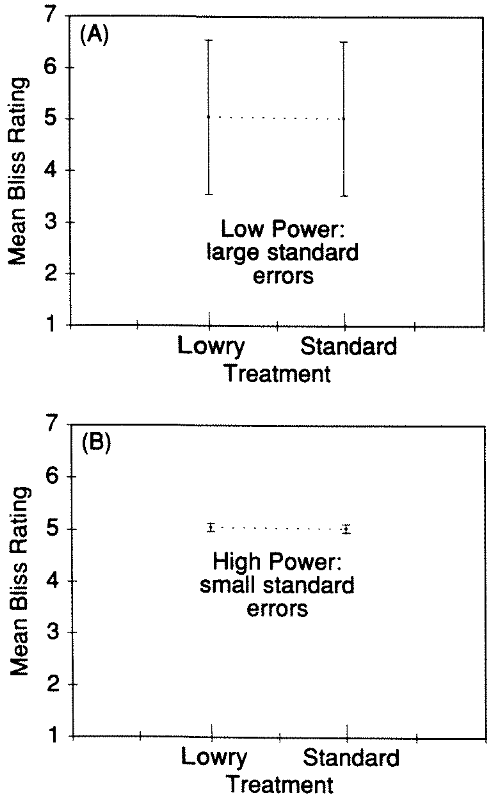

Figure 2. Possible realities corresponding to "no significant difference" with respect to Lowry's data. Top panel (A) depicts low experimental power (large standard error bars); bottom panel (B) depicts high experimental power (small error bars).

The hypothesis-testing procedure that Loeb described in the terse, dense, precise, and scientific-sounding paragraph reproduced above has nothing do with the interesting conclusions. It tells us nothing about what the pattern of population means looks like or how confident we can be about the inferred pattern (i.e., how much statistical power there is). Rather, it merely confirms the banal conclusions, telling us again that it is not true that various sets of population means are identical to one another. In short, the information provided by the PPE procedure, embodied in Figure 1, *subsumes* the standard hypothesis-testing procedure embodied in the quoted paragraph. If you have the paragraph, there is still a need for the figure. But if you have the figure, there is no need for the paragraph. Hypothesis testing is superfluous.

2. Marital Therapy Techniques

Die-hard hypothesis-testing aficionados might argue that Loeb's data, just described, are not typical psychological data. Loeb's experiment involved complex factorial designs, a specific hypothesis about the form of obtained functions (linear), a fancy transformation of the dependent variable; this is not the kind of bread-and-butter experimentation that is so common in our field. What about a simpler experimental design wherein there are just two groups, and all you want to know is whether the groups differ from one another? Surely you do not need a graph. A simple *t* test will do. Or will it?

Two treatments. Let us consider another example (again fictionalized). A clinical psychologist, Jonathan Lowry, developed a new marital-therapy treatment (called the Lowry Treatment). He did an experiment to test the effectiveness of the Lowry Treatment, relative to a much more time-consuming and expensive treatment then in vogue, universally referred to as the Standard Treatment. Forty married couples were randomly assigned one of two groups. Couples in the first group underwent the Standard Treatment, whereas couples in the second group underwent the Lowry Treatment. The outcome measure was the rating of marital bliss (on a 1-7 scale) a year after the treatment. Lowry's hope was that the Lowry Treatment would be just as good as the Standard Treatment, in which case the Lowry Treatment, being simpler and cheaper, would be preferable.

To his delight, Lowry found no difference between the two treatments. He wrote an article about his experiment which he submitted to the premier marital therapy journal, *Eternal Togetherness* (*ET*). He expressed his main finding this way:

> The mean rated bliss of the Standard and Lowry treatment groups were 5.05 and 5.03, respectively. The difference between the groups was not statistically significant, $t(38) = 1.06$, $p > .05$.

The *ET* reviewers thought that the lack of difference between the two treatments had important practical implications, for it meant that the same degree of marital bliss could now be attained much more easily than had previously been possible, and the *ET* editor was thus inclined to publish Lowry's article. The editor was somewhat nervous about publishing a conclusion that relied on accepting the null hypothesis, because it had been firmly drilled into him during his graduate training that accepting the null hypothesis is unacceptable. He thought that at least Lowry should do a power analysis. However, although the editor had never admitted it to anyone, he did not actually understand power very well. After pondering the problem for awhile, he simply accepted the article without changes.

The meaning of "no significant difference". When I read Lowry's article, I was irritated. What did "no significant difference" mean? As I have noted earlier, it *could not* imply that the population means of the two treatment groups are identical. That is a mathematical impossibility. However, *identity* of treatment groups is not really an issue in this practical arena. What is important is this question: Are the two treatments *sufficiently* similar so that one is justified in opting for the easier-to-use, cheaper Lowry Treatment over the Standard Treatment?

The "no significant difference" that Lowry had reported could reflect any of many possibilities. To simplify, consider two polar alternatives. The first is that Lowry was a sloppy researcher, and that there was so much variability within his two treatment groups (i.e., such low experimental power) that the actual population mean difference between the two groups could plausibly be just about anything. This possibility is illustrated in Figure 2A, wherein the small solid circles represent the two group means and the error bars represent the standard errors. Note that the size of the error bars in Figure 2A (large) provide a direct reflection of the power (low). At the risk of redundancy, I emphasize that standard error bars always provide a direct reflection of experimental power: the larger the standard errors, the lower the power.

The second possibility was that there was low variability within the treatment groups (i.e., high experimental power) such that any actual population difference between the two groups would have to be quite small. This possibility is illustrated in Figure 2B. The *ET* editor's intuition was correct: some kind of power analysis should have been done.

In addition to knowing about experimental power, it would be of substantial practical interest to know the *standard deviations*[3] of each of the two treatment groups. Such knowledge would provide some indication of the *range* of martial bliss that any particular troubled couple might expect to achieve given either treatment. For instance, if the standard deviation of the Lowry Treatment group were small, any couple administered this treatment could be assured of eventual bliss fairly close to the mean of about 5; conversely, given a large standard deviation, the precise magnitude of any given Lowry Treatment couple's eventual bliss would be less certain.

In short, Lowry's article provided few clues about anything having to do with the *variability* of marital bliss.

Although information about variability is not directly accessible in Lowry' terse description of his results, it is partially computable from the sample sizes, the sample means, and the *t* value. With this information, I was able to compute that the standard error of the difference between the two population means was about 0.14, which is quite small, given that the entire bliss scale goes from 1 to 7. It appeared that Lowry' experiment had relatively high experimental power; that is, in practical terms, any *actual* difference between the two treatment population means must be of little consequence. Thus, Lowry's actual data were more in accord with the Figure 2B example than with the Figure 2A example.

Although I couldn't compute the individual standard deviations of the two groups, I could compute the mean[4] standard deviation of the two groups, which is 0.434. Insofar as the two groups have similar standard deviations, this tells us that a couple receiving the Lowry Treatment (or the Standard Treatment for that matter) would, with about 95% probability, end up with marital bliss of within about two standard deviations of the mean or, roughly speaking, somewhere between 4 and 6. This is important information for anyone actually considering one of the treatments.

To get more complete information, I e-mailed Lowry, asking him for his raw data. Later the same day, Lowry e-mailed the data back to me. Electronically cutting the numbers from Lowry's e-mail message and pasting them into a previously prepared Microsoft Excel spreadsheet allowed me to immediately calculate everything I wanted to know. What I discovered was interesting and somewhat unexpected: the individual treatment group standard deviations were 0.608 for the Lowry Treatment and 0.086 for the Standard Treatment. Thus, the Standard Treatment, although more costly, is more certain in terms of what a given couple's bliss will actually turn out to be.

To generate a graphical representation of all this information, I pasted the means, standard deviations, and standard errors from Excel directly into my graphing application. With a couple of mouse clicks and keystrokes, I got the graph shown in Figure 3.

In this plot, the two black circles represent the two sample means. Each mean has two error bars associated with it, representing the standard error of the mean (shorter bar) and the standard deviation of the group (longer bar). I assert, as I did with the Loeb example, that this plot conveys the information carried by the standard hypothesis-testing procedure, plus additional, more interesting information. The virtual identity of the two means, in conjunction with the sizes of the standard error bars conveys the hypothesis-testing information that the groups are "not significantly different." That the error bars are relatively small indicates high power, which, in turn, implies that the actual difference between the two population means must also be small. The sizes of the standard deviation bars provides information about the range of where a random couple in either treatment would likely fall given that they had one treatment or the other. In short, this simple figure visually and intuitively conveys all the important and useful information about Lowry's results that took me a couple of paragraphs to convey textually. If Lowry had substituted something akin to Figure 3 for the APA-approved description of his results that appeared in his *ET* article, his readers would have had a much easier time becoming much more informed.

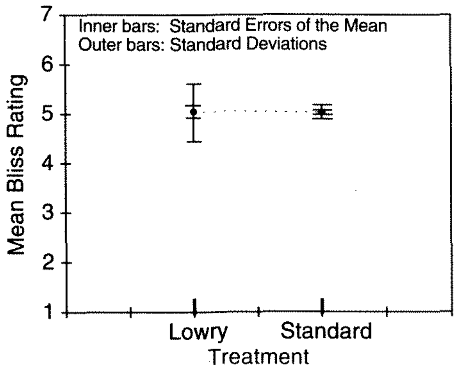

Figure 3. Lowry's data: Mean rated marital bliss for two marital-treatment conditions. Shorter error bars around each mean represent standard errors of the mean. Longer error bars represent estimated population standard deviations.

TODAY'S EASY-TO-USE COMPUTER GRAPHICS

I am by no means the first to argue that graphical representations in general, and the PPE procedure in particular, are useful techniques for understanding and conveying information about the data from some experiment (cf. Tufte, 1983, 1990; Tukey, 1977). It is my hope that the preceding examples, anecdotal though they may be, help illuminate why this is so. In this final section, I will make some comments about the nitty-gritty of actually implementing the kinds of graphical representations shown in Figures 1-3.

We Shouldn't Throw Away our Statistical Packages

There are a multitude of sophisticated and easily obtainable computer applications for doing virtually any conceivable kind of statistical analysis. Even given what I have been saying, I believe these applications to be very useful. I believe, however, that we should view their primary use as summarizing raw data and generating descriptive statistics, such as means, standard deviations, standard errors, and mean square errors. From these applications, we can get the raw material used to generate plots of the sort shown in Figures 1 and 3.

Currently Available Graphing Applications

I would like to make a couple of points about presently available graphing applications. First, I will talk about such applications in general, and then I will describe an informal survey that I have conducted.

Two General Categories of Graphing Applications

Generally speaking, scientists use two different kinds of graphing applications: those that are associated with other applications (e.g, with statistical or spreadsheet applications), and those that are "stand-alone." With rare exceptions, I prefer stand-alone applications for several reasons. First, they tend to be more powerful, more flexible, and easier to use than graphing "features" that are subsidiaries of something else. Second, the across-application cut-and-paste process has become so simple that it makes sense to use each application for its primary function, in conjunction with transferring data from one application to another. Recall my descriptions of my interactions with Dr. Lowry: I originally cut the raw data from one application (my communications application) and pasted it into another (Excel). Then I cut the Excel results and pasted them directly into a third application (my graphing application). All of this took less than a minute.

Who Uses What Applications?

What are the popular graphing applications at present? In the beginning (that is, starting around 1986) the application of overwhelming choice was the CricketGraph, which ran only on the Macintosh. Even PC users somehow ferreted out their friends' Macs (usually in the dead of night) and learned enough about them to use CricketGraph for their graphing chores.

As the years passed, however, CricketGraph fell upon hard times. It was sold first to Xerox, then to Computer Associates. Neither company appeared eager to update the application, and it languished, clearly becoming yesterday's technology. It was generally incompatible with Macintosh System 7, and even caused some Macs running under System 6.x to crash.[5] Within the past year, an update of CricketGraph (Version III) has materialized. In the intervening time, however, a number of disgruntled users turned elsewhere. Two applications in particular—KaleidaGraph and DeltaGraph—proved popular with those renewing their graphing arsenals. My personal favorite is KaleidaGraph, which I used to create Figures 1-3.

While writing this article (in October, 1992), I became curious about what my colleagues used for graphing. Accordingly, I carried out a very informal, nonrandom, and unscientific survey in which I first asked people to identify themselves as Macintosh, DOS, Windows, OS/2, or UNIX users, and then asked what they used to graph their data. I e-mailed this query to all psychologists on my e-mail address list, which included a total of 131 names. Because one of the "names" was MacPsych, the entry into a popular Macintosh chitchat network, the survey recipients were highly biased toward being Macintosh users. Nonetheless, the results are instructive.

Within a couple of days, 94 people had responded, 93 of whom used some graphics application or another.[6] Some of the respondents used more than one application, and 129 total mentions of applications were tallied. Table 1 lists the frequencies with which various applications were mentioned, subdivided by operating system. As anticipated, given the recipient bias, the large majority (87) of the applications mentioned were run under Macintosh. The remaining mentioned applications were run under DOS (25), Windows (5), and UNIX (12). No one reported using OS/2.

In Table 1, under each operating system is listed the frequency with which various applications were reported to be used (many respondents used more than one application). I have collapsed mentions of applications that are not specifically graphing applications under the heading Statistics/Spreadsheet. I counted a statistical or spreadsheet application only if the respondent specifically mentioned using the application's graphing capabilities. Thus, for example, a number of people described doing data manipulation in Excel and shipping the results to CricketGraph for graphing. For such a person, CricketGraph would be counted, but Excel would not. In all, spreadsheet and statistical applications were reasonably popular for graphing, constituting 30% of all mentioned applications. It is of some note that Microsoft Excel was the only spreadsheet mentioned.

Table 1
Frequency and Percentages of Various Mentioned Graphing Applications for Four Different Operating Systems

Application	Frequency	Percent
	Macintosh (n = 87)	
CricketGraph	39	45
DeltaGraph	13	15
KaleidaGraph	10	11
SigmaPlot	1	1
Igor	2	2
Statistics/spreadsheet	22	25
	DOS (n = 25)	
SigmaPlot	7	28
Harvard Graphics	8	32
InPlot	1	4
Fig-P	1	4
Statistics/spreadsheet	8	32
	Windows (n = 5)	
CricketGraph	1	20
Charisma	1	20
Statistics/spreadsheet	3	60
	Unix (n = 12)	
S	5	42
Gnuplot	2	17
Statistics/spreadsheet	5	42

Macintosh applications. Table 1 indicates an obvious winner among Macintosh users: the venerable CricketGraph turned up 45% of the time, with DeltaGraph and KaleidaGraph trailing quite far behind. A new highly flexible application, Igor, was enthusiastically endorsed by two users.

Other applications. DOS users reported being unhappy with the general state of DOS graphing applications. The only ones mentioned by more than one person were SigmaPlot and Harvard Graphics. (Of some interest is that only a single Macintosh user mentioned the reasonably respected SigmaPlot, although there exists a Macintosh version). Many DOS users reported that they used Macintoshes to do their graphing.

There were surprisingly few Windows users. Of the five Windows respondents, one used Charisma, and the other used CricketGraph.

UNIX users generally favored the AT&T application, S.

CONCLUSIONS

The main argument that I have tried to make in this article is that hypothesis testing is the wave of the past (and never should have been a wave at all). Characterizing conclusions in hypothesis-testing terms requires reducing the complex, multidimensional information that generally emerges from an experiment into one or more binary decisions that are almost always logically predetermined to begin with.

I have argued that presenting data in the form of one or more well-designed graphs—particularly graphs that represent the relevant sample means along with various measures of inferred variability—potentially conveys the interesting and important information from the experiment in a manner that (1) is immediate and direct and (2) does not entail a pseudoprecise attribute (viz., $p < .05$) that does little but fool naive readers into thinking that some important conclusion about reality has been made. In particular, the size of the standard errors of the mean provides a direct and intuitive visual measure of how precisely the locations of the relevant population means—and thus the overall pattern of population means—can be inferred.

Given this strategy, it is important to have powerful and easy-to-use tools. There are many such tools in today's microcomputer environment. Any of the applications listed in Table 1 would be perfectly adequate for the task, although obviously the applications differ along a variety of dimensions.

I believe that the family of PPE techniques, illustrated in Figures 1-3, have enormous potential for efficiently conveying information about experimental results. I hope that members of our discipline, like our natural-science brethren, will begin availing themselves of this potential more than is presently the case. In a *Memory & Cognition* editorial (Loftus, 1993) I pursue this hope further, and provide it with more teeth.

REFERENCES

BAKAN, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, **66**, 423-437.

COHEN, J. (1990). Things I have learned (so far). *American Psychologist*, **45**, 1304-1312.

GIGERENZER, G., SWIJTINK, Z., PORTER, T., DASTON, L., BEATTY, J., & KRUGER, L. (1989). *The Empire of chance: How probability changed science and everyday life*. Cambridge, U.K.: Cambridge University Press.

LOFTUS, G. R. (1991). On the tyranny of hypothesis testing in the social sciences. *Contemporary Psychology*, **36**, 102-105.

LOFTUS, G. R. (1993). Editorial comment. *Memory & Cognition*, **21**, 1-3.

MELTON, A. W. (1962). Editorial. *Journal of Experimental Psychology*, **64**, 553-557.

NUNNALLY, J. (1960). The place of statistics in psychology. *Educational & Psychological Measurement*, **20**, 641-650.

TUFTE, E. R. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.

TUFTE, E. R. (1990). *Envisioning information*. Cheshire, CT: Graphics Press.

TUKEY, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

NOTES

[1] The basic idea is as follows. The null hypothesis typically states that some population parameter (e.g, a population mean, the difference between two population means, a population correlation) is identically equal to some constant. Only rarely could such a hypothesis actually be true. Consider, for example, an experiment in which two treatments, A and B, were being compared. The null hypothesis would be "the difference between the Treatment A outcome score and the Treatment B outcome score is zero (to an infinite number of decimal places)." Such a null hypothesis could not literally be true. So the results of a significance test do not, as advertised, tell us whether or not the null hypothesis is actually false (we know a priori that it is false). Rather, the results simply tell us whether there is sufficient experimental power to detect the population mean difference that inevitably exists.

[2] For illustrative purposes, I assume throughout this article that sample means are the primary data of interest. All arguments could be equally well applied to any sample statistic.

[3] The term "treatment group standard deviation" carries with it some ambiguity: it could refer either to the group's actual standard deviation, or to the estimate of the relevant population standard deviation (these two statistics differ by a factor of $n/[n-1]$). For the purposes of this discussion, I refer to the latter.

[4] Not the arithmetic mean, actually, but the standard deviation of the mean of the two individual treatment-group variances.

[5] Its compatibility with Version 6.x turns out to depend on exactly what ROM the computer has. We determined this factoid by running CricketGraph 1.3 on two seemingly identical Mac IIci computers with the same floppy-based system. It worked on one and crashed the other. The only difference between the two computers was the ROMs they used.

[6] One respondent, a world-famous visual perception expert, claimed to still do his graphing by hand.