Exactly how many p values is a picture worth?

A commentary on Loftus's plot-plus-error-bar approach

G. ROLFE MORRISON and BRUCE WEAVER
McMaster University, Hamilton, Ontario, Canada

Loftus (1993b) has recently argued that hypothesis testing is largely irrelevant and should be replaced with a "plot-plus-error-bar" (PPE) presentation of the data. We agree that such figures can be useful. However, they do not preclude standard hypothesis testing procedures, and they should be used to supplement rather than to supplant the latter. The insufficiency of PPEs is most apparent in the case of mixed designs and factorial within-subject designs. In such cases, there is no single value that is appropriate to represent the standard error.

The oft' flogged horse of null hypothesis testing has managed to survive many whippings over the years, but apparently it is still subject to maltreatment. In a recent article in this journal, Loftus (1993b) proposed that familiar statistics such as t , F , U , and so forth, be largely if not wholly supplanted by the "plot-plus-error-bar" (PPE). In his view, PPE "provides essentially all the information provided by a hypothesis-testing procedure, plus additional information" (p. 250). It is suggested that such plots have not been used because, "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..." (p. 251). This may have been the case in the quite recent past, but not in the preelectronic age. Possibly the author was never faced with doing a complex analysis of variance on an electromechanical calculator.

Loftus (1993b) uses two (fictitious) examples to illustrate his points, and then discusses the "easy-to-use" graphics and statistical applications available to today's researcher. With respect to the latter, he advises against discarding our statistical packages, because they are useful in generating the type of plots he advocates. Loftus's proposals are not to be taken lightly, for in a recent editorial (Loftus, 1993a) he states that, as Editor-elect of *Memory & Cognition*, he intends to encourage the use of PPE in place of traditional hypothesis-testing procedures. Regrettably, the arguments Loftus makes for his position are somewhat less than convincing. We will address in turn the two examples he uses (Loftus 1993b) and then conclude with some comments on his advo-

cacy of statistical software packages as useful tools for generating PPEs.

EXAMPLE 1: LOEB'S DATA

The first example uses fictional data from a fictional psychologist.

Loftus's (1993b) Figure 1 is redrawn here as our Figures 1A and 1B, and for our purposes is self-explanatory. It is a PPE plot of the data from a $2 \times 2 \times 8$ within-subject experiment. To quote again, "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" (Loftus, 1993b, p. 251). Loftus's point is that this plot supplies all the information that the reader requires and that it was ridiculous for the journal editor to ask for F values as well.

One might be tempted to agree except that, as noted, this is a completely *within*-subject design, which raises the question of how the error bars were obtained. Loftus (1993b) gives no indication of how this is to be done. In Loftus (1993a), however, footnote 5 refers one to Loftus and Loftus (1988, pp. 411–413 and 426–429) for a discussion of appropriate error bars in repeated measures designs. Actually, these pages are of little help for the problem at hand, because the discussion is limited to one-way designs. However, they are worth commenting on in the present general context because of their treatment of independent as well as repeated measures designs.

For independent measures, the standard error of each mean is to be estimated by the $\sqrt{(MS_W/n_i)}$, and for a (one-way) repeated measures design, it is to be estimated by $\sqrt{(MS_I/n)}$, where MS_W and MS_I are the within and the treatment $\times$ subject mean squares, respectively, and n_i and n are the number of measures upon which the mean is based. Note that with the repeated measures design, and in the common case of the independent mea-

We thank Lorraine Allan for useful comments on a draft of this article. Requests for reprints should be directed to G. R. Morrison, Department of Psychology, McMaster University, Hamilton, Ontario, Canada, L8S 4K1 (e-mail: morrison@mcmaster.ca).

Figure 1. (A and B) Redrawn from Figure 1 of Loftus (1993b): "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." (C) The data of panel B redrawn with the standard error increasing with the mean. The *average* standard error is the same as that in panel B.

asures where all n_i are equal, all means in the set will have the same estimated standard error.

The Equality of Standard Errors

We have all been taught that to do an analysis of variance (ANOVA) the assumption of homogeneity of variance is required. That is, one must assume that the population (error) variances are equal for all treatments. We also know that this assumption is seldom if ever exactly true in practice, and that it can be violated without grave consequences, particularly if all n_i s are equal.

However, if we are going to do an ANOVA and assume homogeneity of variance, it could be argued that this assumption should be carried through and a common estimate of the standard error be used for each mean. If, on the other hand, we follow Loftus's direction and

abjure the ANOVA and the F ratio, then there is no necessity to make the equal variance assumption and every reason not to. Standard error estimates can and should be calculated separately for each mean, using the within-cell variances in the independent measures design and the residual within-cell variances (after removing the between-subject effects) in the repeated measures design. Indeed, this course of action is recommended by Loftus and Masson (1994).

This, although useful when the homogeneity of variance or sphericity assumptions are violated, does not address the major problem. In cases other than those of independent measures and the one-way repeated measures design, there will typically be more than one error term associated with each mean; for example, in the "mixed A×B" design, there will be both a within-subject and a between-subject error term associated with

each mean, and in a two-way repeated measures design, there are three error terms, and so forth.

In the three-way repeated measures design under discussion there are three main effects, three double interactions, and a triple interaction, and associated with each effect is its own error term. There is no requirement or expectation that these error terms be equal. On the contrary, they are expected to be estimates of different nonzero sources of variation. Except in rare circumstances, there is no rationale for pooling them. In comparing any two means, the appropriate standard errors, or the standard error of the difference, will depend on which two means are compared. There is no way that a single value can represent the standard error of the mean in this case. Inspection of Figures 1A and 1B, however, reveals that the plotted error bars are all equal, which suggests the use of a common error term. The use of error bars in this case—especially equal ones—is both meaningless and misleading.

Loftus and Masson (1994) recognize this problem and suggest, but do not discuss in any detail, several possibilities. Their first suggestion is that “in such cases, one should adopt the strategy of plotting confidence intervals that can be used to assess patterns of means or contrasts that are of greatest interest” (p. 485).

It is not clear how “greatest interest,” or conversely “little interest,” is to be defined. One way might be in terms of the perceived theoretical interest. This, however, may well be different for the author and the reader. Alternatively, a possible effect may be of little interest, because there is only a trivial difference between a pair of means. If this is how interest is to be defined, then there is a problem, since effects of little interest can be demonstrated to be of little interest only by plotting them or by doing the conventional significance tests.

A second suggestion is that “one might even plot more than one confidence interval for some means, or construct more than one plot of the data” (p. 485).

In, for example, a two-way mixed design, it might be possible that two intervals can be plotted around each mean, one based on the within-subject and the other on the between-subject error, without making the figure too confusing to be meaningful, but it is doubtful if the practice could be extended to any more complex case. In the above example, where there could be seven confidence intervals for each mean, the result would be chaotic, even more so if the sphericity assumption was violated for one or more of the terms.

Loeb's Data With Unequal Variances

Since Loftus, as noted above, felt free to select the results to demonstrate his points, we will do the same. Let us assume that Loeb's experiment had consisted only of the conditions shown in Figure 1B, the “no-verbal-encoding” conditions. We will assume further that all conditions were “between subjects” so that the individual standard errors could be calculated. Let the means be as plotted.

Loftus maintains that a simple inspection of the figure clearly indicates that higher uncertainty leads to poorer performance. But does it? Would it not be reasonable to suggest that the variance might well increase with the mean? In other words, the magnitudes of the standard errors would increase with increasing exposure duration and there would be considerable overlap of the error bars throughout the range, as shown in Figure 1c, not just at the lower durations. If that was the case, might one not be led to wonder if the true functions were not, in fact, parallel? Might not two readers seeing the same data be led to conflicting conclusions?

We would argue that means certainly should be plotted with error bars if the appropriate ones were obtainable, or without if not, but *never* with misleading ones. In any case, such plots must be supplemented with an appropriate inferential statistic. In short, we would happily see what we believe Loftus wants us to see in these data if only a line or two of text were added stating, for example, that 95% of the exposure $\times$ uncertainty interaction could be accounted for by the linear component and that the t value for this component had a probability of $< .01$. In this case, one p value would be worth a thousand pictures.

EXAMPLE 2: LOWRY'S DATA

We turn now to Loftus's (1993b) second example, with the results shown in Figure 2, which is Loftus's Figure 3 (p. 254). In this case, we have two groups of subjects from an experiment by Lowry (again a fictitious name), one group subjected to a “standard treatment” and the other to the “Lowry treatment.” Lowry wished to

Figure 2. Redrawn from Figure 3 of Loftus (1993b). “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.”

show that the latter was as effective as the former. Lowry reported that the treatment means were 5.05 and 5.03, and that the difference between the groups was not statistically significant [$t(38) = 1.06, p > .05$]. Loftus maintains that Lowry's conclusions are misleading because the information that there is a large difference in the standard deviations of the two groups is not reported (true, but see below) and that his (Loftus's) treatment of the data gives an accurate picture of what has taken place (perhaps less true than he thinks).

In this example, Loftus (1993b) wishes us to focus on the differences in standard deviations between the two groups. But given the whole thrust of his article, there is a surprising omission in the figure, namely, there are no standard errors given for the *standard deviations*. True, the standard deviations *look* quite different, but looks can depend on how things are plotted. How much bigger must one sample standard deviation be than another before one can conclude that there is a meaningful difference in the population standard deviations?

In this case, one can make some inference, because Loftus happened to include some (but not all) of the relevant information in the text. Taking Lowry's figures, one can transform the standard deviations into variances S_1^2 and S_2^2 . Under the null hypothesis of equal variance (and assuming reasonably normal distributions), S_1^2/S_2^2 (or S_2^2/S_1^2) will be distributed as F . For these data, the observed F ratio is about 50. We are told that there were 40 subjects in all, so, assuming equal sample sizes of 20 and 20, it would be concluded that there was a significant and substantial difference in the two population variances. It should not be left to the reader to do these calculations since one line added to the text would be all that would be required. Again, the figure is useful, but not sufficient.

It might be noted that, in the first example, all variances are to be *assumed* to be equal. In the second example, the variances for two treatments are shown as unequal. Are we to conclude from this that we are free to choose which of these best demonstrates what we would like to demonstrate?

A Computational Error, But Where?

Before leaving the above, we must note a serious inconsistency in the reported data. As stated, the means are reported to be 5.05 and 5.03 and the t value to be 1.06. This would seem to imply that the standard error of the difference between means is $(5.05 - 5.03)/1.06 = .0189$. Loftus claims, however, that he computes the standard error of the difference to be .14. This value of .14 is consistent with the standard deviations computed by his computer to be .086 and .608 (assuming equal sample sizes of 20), but highly inconsistent with the reported t value. There is an error somewhere, but where? We will return to this below when we discuss the use of computer packages.

"What" Versus "How" Data Are Reported

Regardless of the above, the problem in this second example is not so much with how the data are reported,

but rather with what is reported. In this case, we would expect that the author would note the marked difference in variability (if, in fact, this is the case) and comment on it. A line of text or a small table giving the values of the respective standard deviations or variances and an F ratio would be more meaningful to the reader than Loftus's plot.

Suppose for a moment that Lowry's data were such that both the means and standard deviations were quite similar to each other but that the two groups differed markedly in skewness. This is yet another difference that would not be indicated in the plot recommended by Loftus, but could easily be mentioned in a description of the data.

STATISTICAL SOFTWARE PACKAGES

We now turn to Loftus's comments on the dangers of entrusting your data to computer programs that then generate a z , t , or F value. He cites Cohen (1990) as stating that some peddlers of statistical software packages claim, correctly, that to use them you do not have to understand statistics. We fully agree with Loftus and Cohen on this point. Later in the article, however, Loftus suggests that these programs are useful for generating the kind of descriptive statistics and data plots that he advocates without passing on the caveat that unless you know what you are doing *and* what the program is doing, it can make as big a mess of descriptive statistics as of inferential ones. (A simple example is that one can throw a set of numbers into a program and ask for and get a standard deviation. However, is that "standard deviation" calculated by using n or $n-1$ as the divisor? Without knowing the program or at least consulting the manual, you have no way of knowing.)

In the case of Lowry's experiment, Loftus (1993b) states: "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" (p. 254). As noted already, there seems to be something seriously wrong in these calculations.

A more complex example comes from Loftus's (1993b) Figure 1. Note that the figure caption states that "solid lines represent best-fitting linear functions." We are not given a definition of "best-fitting" and must guess at it. The chances are very high that what was used was the least squares procedure, especially if a standard statistical program was utilized.

The least squares procedure may qualify as giving a "best fit" line only under certain limited circumstances. In this case, the function of theoretical interest is an exponential one, $p = (1.0 - e^{-d/c})$. However, rather than plot this, a log transform [$P = -\ln(1-p)$] is applied so that P becomes a linear function of d (duration). Thus, it is probable that the least squares criterion applies to the transformed data. Procedures that require the transformation of the dependent variable often produce rather unsatisfactory descriptions of the untransformed scores.

One requirement for the least squares procedure to be a "best fit" one is that the error variance be constant throughout the range (or at least not change systematically). But is the error variance of the log scores or the original scores constant, or neither? This would seem to be the type of question that must be considered before entrusting the data to a computer package even for descriptive statistics. This is especially true in this case, as Loftus (1993b) states: "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* [italics added]" (p. 251).

If these exact relationships are important, then it is crucial to consider the most appropriate curve-fitting procedure and not to leave it to the whim of a computer program. It is quite possible that Loeb did in fact do this, but Loftus makes no mention of it and does not seem to consider it to be of importance. For a more extended discussion and example of how the wrong program can distort descriptive analyses, see Morrison, Morrison, and Keating (1992).

WHY STANDARD ERRORS?

Loftus wants us to plot the mean plus or minus the standard error. We ask somewhat rhetorically, why the standard error? If the purpose is simply to indicate graphically some measure of the variability of the mean, then would not some other measure such as the standard deviation, the mean deviation of the interquartile range, and so forth, be equally suitable? The answer, we suspect, is that the idea is not simply to give a descriptive measure of variability, but to give a confidence interval.

If this is the case, two questions arise. First, is the PPE not coming dangerously close to hypothesis testing? Second, why are we to plot the mean plus or minus one standard error? If we are talking about a normal distribution of sample means, the probability of the true mean falling in the range of sample mean plus or minus one standard error is about .67. If we are to use the PPE procedure, should we not at least plot the mean plus or minus *two* standard errors?

A criticism that is sometimes leveled at the use of p values is that there is sometimes the mistaken implication that a small p value must indicate a large effect, overlooking the fact that with a false null hypothesis p may be made as small as you wish by using a large

enough sample. The same, however, holds true for the standard error and, we believe, may be even more subject to misinterpretation. Would it not be better, in order to achieve what appears to be Loftus's goal, to plot a measure of variability that is independent of the sample size?

CONCLUSIONS

To sum up, here are our points of agreement and disagreement with Loftus. First, we agree with Loftus that the conventional hypothesis-testing procedures often leave much to be desired. However, we do not agree that they can be adequately replaced by the PPE procedure he proposes. Second, we agree that graphical plots of data are a good thing and that, when they can be calculated appropriately, error bars may be useful. These plots, however, seldom supplant the appropriate inferential statistics and should be used with, rather than instead of, the latter. Inappropriate and incorrect error bars are misleading and to be avoided. Third, we agree that the use of computer programs to compute inferential statistics without understanding the statistic or the program is dangerous, but would add that this is also true for descriptive statistics. Finally, we agree that programs to plot results are a boon to the researcher, but, like other programs, are not in themselves an adequate substitute for thought.

REFERENCES

- COHEN, J. (1990). Things I have learned (so far). *American Psychologist*, **45**, 1304-1312.
- LOFTUS, G. R. (1993a). Editorial comment. *Memory & Cognition*, **21**, 1-3.
- LOFTUS, G. R. (1993b). A picture is worth a thousand p values: On the irrelevance of hypothesis testing in the microcomputer age. *Behavior Research Methods, Instruments, & Computers*, **25**, 250-256.
- LOFTUS, G. R., & LOFTUS, E. F. (1988). *Essence of statistics* (2nd ed.). New York: Random House.
- LOFTUS, G. R., & MASSON, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, **1**, 476-490.
- MORRISON, G. R., MORRISON, S. R., & KEATING, D. P. (1992). On estimating processing variance: Commentary and reanalysis of Kail's "Developmental functions for speeds of cognitive processes." *Journal of Experimental Child Psychology*, **54**, 288-307.

(Manuscript received January 3, 1994;
revision accepted for publication April 1, 1994.)