Behavior Research Methods Instrumentation and Computers
1995 , 27, 57-59

Data Analysis as Insight: Reply to Morrison and Weaver

GEOFFREY R. LOFTUS
University of Washington

I begin by underscoring a fundamental point of agreement between Morrison and Weaver and me: that data analysis should entail designing whatever set of techniques optimizes insight into whatever question the experiment was designed to address, rather than implementing some rote set of pre-ordained rules and regulations. Against this shared philosophical backdrop I then (1) reiterate problems with hypothesis testing and (2) address some of the quite pertinent issues that Morrison and Weaver raise with respect to computation of confidence intervals.

Morrison and Weaver (1994) end their commentary by noting that "...programs to plot results...are not in themselves an adequate substitute for thought" (manuscript p. 13). Although this assertion sounds like a truism, it is not. Rather it is a legitimate response to one aspect of a trend about which numerous psychologists (including me) have become increasingly concerned over the years—that of data analysis as no-brainer. I believe, as I have asserted elsewhere, that this trend is largely rooted in an almost exclusive reliance on hypothesis-testing procedures which, as they are modally used, amount to reducing a rich and complex data set to an impoverished and uninformative string of significant/nonsignificant decisions.

Throughout Morrison and Weaver's remarks is woven the theme that such gratuitous information reduction is unacceptable—that data analysis should be carried out and reported in such a way as to optimally exploit the information carried by a data set. For example, they note that if confidence intervals are reported, there is no need to make the standard homogeneity of variance assumption; one can individualize confidence intervals for individual conditions. Likewise, they note that if a critical conclusion depends on an over-conditions comparison of *variances*, then confidence intervals of the variances should be reported; that if infrequently reported statistics (like skewness) provide insight about the question being addressed then such statistics should be computed and reported. And so on.

In short, there is a fundamental basis of agreement between Morrison and Weaver and me: that data analysis should entail designing whatever set of techniques optimizes insight into whatever question the experiment was designed to address, rather than implementing some rote set of pre-ordained rules and regulations (e.g., "do the appropriate ANOVA and leave it at that.") To the degree that the general attitude embraced by Morrison and Weaver enters the methodological zeitgeist, they, and I, and many others will rejoice. Given our shared attitudes, our points of disagreement seem relatively unimportant.

With this foundation in mind, let me turn to some specific points of agreement and disagreement.

On the Value of Hypothesis Testing

As I have emphasized, Morrison and Weaver and I espouse the same fundamental goal, which might be expressed as "creative and flexible data analysis as a route to optimal insight." Morrison and Weaver assert that the

The writing of this article was supported by an NIMH grant to the author. Correspondence concerning it may be sent to Geoffrey R. Loftus, Department of Psychology, University of Washington, Seattle, WA 98195; email: gloftus@u.washington.edu.

plot-plus-error-bar (PPE) method favored by Loftus (1993a; see also, Loftus, 1991, 1993b, 1993c) is burdened with various difficulties that ought to render it subsidiary to standard hypothesis-testing procedures. Below, I will address some of Morrison and Weaver's complaints about the PPE method. First however, I would like to reiterate two remarks I have made in the past about fundamental problems with hypothesis testing.

Implausible Null Hypotheses

A null hypothesis is a specific statement about relations among a set of population means (e.g, that all population means are identical to one another). In almost all cases, a null hypothesis (at least in the social sciences) can be assumed a priori to be false (e.g., no set of population means are identically equal to an infinite number of decimal places). Thus a test of the null hypothesis, contrary to conventional wisdom, does not usually test the null hypothesis. What it usually *does* test is whether there is sufficient statistical power to detect whatever violation of the null hypothesis must exist.

As an example of this issue, consider Morrison and Weaver's analysis of linear functions (manuscript p. 8). Morrison and Weaver state that "...we would happily [believe in the validity of fundamentally linear functions] if only a line or two of text were added stating, for example, that 95% of the exposure by-uncertainty interaction can be accounted for by the linear component and that the *t* value for this component has a probability of < .01. In this case, one *p* value would be worth a thousand pictures."

There are two quite separate suggestions here. The first—that planned comparisons should be used as a basis for describing the percent variance accounted for by some effect—is, in my opinion an excellent one; as I have stated elsewhere (Loftus, 1993c) I believe that planned comparisons are puzzlingly underutilized in the social sciences. However what about reporting that $p < .01$? This small *p* value asserts that, given the null hypothesis of *zero* correlation between a set of linear weights and a corresponding set of population means, the observed degree of linearity has a probability of less than 1%. But the assumption of an exactly zero correlation is entirely implausible to begin with. So what's the point of considering anything that depends on its validity? It's like saying that the Moon's observed color is highly unlikely given that the Moon is made of green cheese. True, but so what?

Miscast Objectivity

Morrison and Weaver (manuscript p. 7) consider assessment of two not-quite-parallel observed linear functions described by Loftus (1993a). They observe that the functions themselves in conjunction with the error bars around the sample means strongly suggest the underlying (population) functions to be likewise non-parallel. However, Morrison and Weaver describe a plausible scenario (of variances covarying with means) by which the underlying functions might indeed be parallel. They ask "Might not two readers seeing the same data be led to conflicting conclusions?"

Well...yes. And indeed it is the lot of many data sets reported in the social sciences to be similarly ambiguous. Morrison and Weaver seem to imply, however, that a simple ANOVA would clear up this ambiguity, as the ANOVA would cleanly result in a "reject" or a "don't reject" decision.

There are several interrelated problems here. The first, noted above, is that *rejecting* a null hypothesis (in this case, a null hypothesis of parallel curves) generally doesn't tell you anything you didn't know before, as the null hypothesis can almost always be assumed a priori to be false. However, what if the ANOVA results in a failure to reject the null hypothesis? As we have all been taught, this is essentially a "no-decision" outcome. By convention, we couldn't reject the null hypothesis; but nor, by logic, could we infer it to be true. The problem, however, is that many investigators would (either explicitly or implicitly) accept the null hypothesis anyway. Whether we like it or not, a statistical analysis technique has sociological consequences. A consequence of using hypothesis testing is that most humans—even humans who are sophisticated scientists, schooled in statistical reasoning—can't seem to accept a "no decision" state, succumbing instead to the irresistible temptation to divide effects into those that exist (when statistical significance

has been found) and those that don't (when it's not).

Computing Appropriate Confidence Intervals

Morrison and Weaver have performed a very useful service in underscoring a number of oversimplifications in the computation of confidence intervals as described by Loftus (1993a). In general, these problems have been addressed (although certainly not entirely solved) in a forthcoming article that Morrison and Weaver acknowledge (Loftus & Masson, under revision). Here I briefly describe some of Loftus and Masson's points that are relevant to Morrison and Weaver's arguments.

Repeated-Measures Designs

Loftus (1993a) restricted his remarks to entirely between-subjects designs wherein computation of confidence intervals is straightforward and is embodied in the equation,

$$CI = M_j \pm \sqrt{\frac{MS_W}{n_j}} \text{ [criterion } t(df_W)]$$

where M_j is the mean of Condition j, n_j is the number of observations in Condition j, MS_W is the (pooled) mean square within conditions, and df_W is degrees of freedom within.

Loftus and Masson argue that any kind of statistical analysis (e.g., ANOVA or PPE) is fundamentally used to elucidate the underlying *pattern* of population means. This reasoning allowed them to propose a confidence interval in an entirely within-subjects design; here MS_{SxC}, the mean square due to subject-by-condition interaction is substituted for MS_W, and accordingly, the formula becomes,

$$CI = M_j \pm \sqrt{\frac{MS_{SxC}}{n}} \text{ [criterion } t(df_{SxC})]$$

where n is now the total number of subjects. Loftus and Masson show that this confidence interval, while not appropriate for inferring the value of any *single* population mean is, like its between-subjects counterpart, appropriate for inferring how much faith one can put in the *pattern* of sample means as a reflection of the underlying pattern of population means.

Homogeneity of Variance

Morrison and Weaver assert that the homogeneity of variance assumption is rarely correct. Although they do not say so, this is an argument for abandoning the ANOVA; however heterogeneity of variance poses no problem for computation of confidence intervals. In a between-subjects design, the confidence interval around M_j becomes

$$CI_j = M_j \pm \sqrt{\frac{MS_{Wj}}{n_j}} \text{ [criterion } t(n_j - 1)]$$

where MS_{Wj} is the mean-square error computed within Condition j only. Loftus and Masson demonstrate an analogous confidence interval for within-subjects designs which is,

$$CI_j = \sqrt{\frac{\text{estimator}_j}{n}} \text{ [criterion } t\,(n - 1)]$$

where n is the total number of subjects, and "estimator_j" is

$$\left(\frac{J}{J-1}\right)\left(MS'_{Wj} - \frac{MS_{SxC}}{J}\right).$$

Here, J is the number of conditions, and MS'_{Wj} is the estimated variance within Condition j after overall subject variation has been removed (see Loftus and Masson for details).

Multifactor Within-Subjects Designs

Morrison and Weaver point out that in multifactor completely within-subjects designs, there are numerous subject-by-effect error terms, and hence numerous possible confidence intervals. Morrison and Weaver acknowledge Loftus and Masson's point that if the error terms are sufficiently similar, they can be pooled—that is, the design can be treated as a one-way design (with JxKxLx... conditions) and a single confidence interval, based $MS_{SxABC...}$, can be computed with little loss of validity. If the various error terms are quite different from one another then computation of a single confidence interval would be, as Morrison and Weaver stress, a misleading procedure. Note, of course, that in such an event the error-term differences would likely themselves become a phenomenon of some in-

terest. (Why, the investigator might ask for example, would subjects be less consistent over levels of Factor 1 than over levels of Factor 2?)

Mixed Designs

A mixed design is more complicated as there would be no reason to suppose that the within- and between-subjects error terms would be the same. Again, Loftus and Masson suggest several means of addressing this problem. First, if appropriate, one could compute whatever confidence interval is most germane to the critical conclusion. Second, one could compute multiple confidence intervals corresponding to the various effects. Third, using some a priori model (e.g., a linear scanning model applied to a Sternberg short-term memory scanning task) one can sometimes reduce a mixed design (e.g., set size within subjects x stimulus type between subjects with RT as the dependent variable) to a pure design (e.g., a between-subjects design entailing different stimulus types and *slope* as the dependent variable).

Et Alia

Finally, a couple of minor miscellaneous comments:

PPE is not Close to Hypothesis Testing

In considering the arbitrariness of choosing to use, say, a standard error versus a 95% confidence interval, Morrison and Weaver pose the question "...is the PPE not coming dangerously close to hypothesis testing?" (manuscript, p. 12) . The answer is: "No." Yes, there is a correspondence between the alpha level used in hypothesis testing and a confidence interval's probability level. But, for reasons I have alluded to above and elsewhere, hypothesis testing and confidence intervals entail different kinds of basic logic (the former testing the plausibility of a data set given some null hypothesis, and the latter directly describing a pattern of population parameters).

Mea Culpa

Morrison and Weaver are correct (manuscript p. 9). I made a computational error in describing the fictional Lowry data. The t value should have been reported as simply, "$t < 1.0$."

References

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

Loftus, G.R. (1993a). A picture is worth a thousand *p*-values: On the irrelevance of hypothesis testing in the computer age. *Behavior Research Methods, Instrumentation and Computers, 25*, 250-256 .

Loftus, G.R. (1993b). Editorial Comment. *Memory & Cognition, 21*, 1-3.

Loftus, G.R. (1993c). On the overreliance on hypothesis testing in the social sciences. Presented at the Psychonomic Society meetings, Washington, DC.

Loftus, G.R. and Masson, M.E.J. (1994). A proposed confidence interval for use in within-subject designs. (Under revision for *Psychonomic Bulletin & Review*).

Morrison, G.R. & Weaver, B. (1994). Exactly how many p values is a picture worth? A commentary on Loftus' plot-plus-error-bar approach. *Behavior Research Methods, Instrumentation & Computers.* (in press).