Bulletin of the Psychonomic Society
1975, Vol. 6 (6), 608-610

Stimuli and subjects in one-tailed tests

JONATHAN BARON
University of Pennsylvania, Philadelphia, Pennsylvania 19174

It is argued that in many experiments in which two kinds of stimuli are compared, the best statistical test is a one-tailed test across stimuli. This view implies that Clark's (1973) argument for testing across stimuli *and* subjects may not apply for some common types of experiments, including his own examples.

Many experiments involve comparing some sort of performance measure for two different kinds of stimuli. These experiments typically involve presenting several examples of each kind of stimulus to each of several subjects. A good example of such a study is that of Rubenstein, Lewis, and Rubenstein (1971). One of the purposes of this study was to find out whether a phonemic representation of a printed stimulus is used in deciding whether or not that stimulus is a word. To test this hypothesis, Rubenstein et al. picked several words and several nonwords, and presented these to subjects one at a time in a word-nonword decision task, timing the response to each stimulus. These were two types of nonwords, homophonic nonwords such as BRUME and nonhomophonic nonwords such as FRUMP. Let us assume that the only systematic difference between these two sets of nonwords is this distinction between homophonic and nonhomophonic. If this assumption is true, the hypothesis of interest can be tested by comparnig the time to reject homophonic and nonhomophonic nonwords; homophonic nonwords should be harder to reject because their sounds alone might tempt the subjects to say that they were words.

Rubenstein et al. tested for an effect of homophoney by testing across subjects, that is, by comparing the variance due to treatments to the variance due to subject-treatment interactions. Clark (1973) has rightly pointed out that this test was not appropriate, because it could turn out significant simply because the homophonic stimuli happened to be harder than the nonhomophonic stimuli, because of the particular sample of each type that was chosen. Clark argued that one needs to test across stimuli as well as subjects; that is, we should take into account the variance due to stimuli within treatments. Clark argues that in experiments of this sort, it is necessary to test across both stimuli and subjects so that we can "generalize a finding to two populations at the same time." His paper is concerned with the details of how to do this.

But just what does it mean to *generalize a finding to a population*? Certainly this phrase does not mean that we assert that the effect will hold for every member of that population. Nor does it mean we assert that that replication of the experiment with a different sample from that population will yield a significant result. We may be willing to bet that the result of a replication is more likely to come out in the direction it did the first time than the opposite direction, but we could bet this without doing any statistical tests at all. It seems to me that the purpose of statistics is only to "generalize to a population" in the loosest sense. What this phrase really means—as Clark implicitly acknowledges—is to rule out alternative interpretations of an effect by showing that the results are improbable if the alternatives are assumed. Viewed in this way, each statistical test we do should correspond to some set of alternative interpretations that we are trying to rule out.

Consider again the Rubenstein et al. result that the homophonic items take longer to reject than the nonhomophonic items. The favored interpretation of this result is that the sound of the item is used in deciding whether or not "it" is a word. But there is an alternative interpretation that must be considered. This is the hypothesis that the result is actually due to stimulus variance, to random variation in the difficulty of the stimuli, irrespective of which treatment group they are in, leading to a mean on one side of zero (no effect) rather than another. This alternative interpretation may be called the null hypothesis for stimuli.

To rule out *this* null hypothesis, or alternative interpretation, the appropriate test would seem to be a one-tailed t test across stimuli (essentially Clark's F_2, but one-tailed). The reason for using a one-tailed test is that the null hypothesis is in fact that the effect is either zero or reversed from what is predicted. The effect would be reversed if there were some (admittedly hard to imagine) mechanism that made homophonic nonwords easier to reject instead of harder. The existence of such a mechanism would have no bearing (we will assume) on the hypothesis of interest. Let us assume that we do this test and it is significant. We have ruled out the null hypothesis for stimuli.

Now the question arises whether there is some other null hypothesis that must be ruled out. Is there, as Clark would argue, a comparable null hypothesis for subjects? Can we describe this other null hypothesis? Well, let us begin by simply substituting into the language we used before, as closely as possible, which now becomes: The result is actually due to subject variance, to random variation in the magnitude of the effect for different subjects, leading to a mean of zero on one side rather than another. Is this disturbing? Imagine the extreme case in which, say, four subjects showed a 100-msec difference in the predicted direction and two subjects showed a 100-msec difference in the opposite direction. In this case, the result would not be significant across subjects, but it could be significant across stimuli. It seems to me that such a result need not disturb our conclusion that our hypothesis is true on the basis of a one-tailed test across stimuli. This finding of nonsignificance across subjects does indeed suggest that the mechanism of interest is absent in some subjects and that there might even be some mechanism leading to a negative effect. But can we argue that the mechanism of interest is present in no subjects? I do not see how we can. The mechanism of interest still exists; high subject variance in its effect is an additional fact, not an alternative explanation.

Perhaps the problem is that we have not found the right statement of the alternative hypothesis. Clark (1973, p. 338, Equation 4b) suggests another. This is that there is in fact no effect of the treatment, no variance resulting from the difference between homophonic and nonhomophonic items, but there is variance due to treatment by subject interactions. But this variance is exactly what we spoke of in the last paragraph; it is based on the idea that subjects differ in the magnitude of the effect, some showing no or a reversed effect and some showing a positive effect. Again the answer is that a significant result across stimuli implies that some subjects show a positive effect, which in turn means that the mechanism of interest exists.

But perhaps the alternative hypothesis does not have anything to do with variances due to different sources. Perhaps it is something more like: The effect occurs for *too few* subjects for it to be of interest. What could this mean? Surely it does not mean that the effect occurs for so few subjects that we cannot predict that a replication with different subjects would yield significant results, since we cannot predict this even from a significant result across subjects. Nor could it mean that we should not put our money, at even odds, on the result coming out in the same direction in a replication, for we do not need a significant result for this bet to be rational. Still another interpretation is that a significant result across subjects means that we can take seriously any failure to get the effect in a different population, say, deaf readers. But this is not true either; to take such a failure seriously, we must test for the presence of an interaction of Groups by Treatments; a significant result next to an insignificant result does not make an interaction. A final interpretation is that there is some critical level of the odds on a significant replication, between zero and a sure bet, that corresponds to a significant result. While this is the case (given an exact replication and ignoring a priori probability), this is not how significance levels are chosen. Rather, as argued, significance levels correspond to a certain probability of the results, given a certain alternative interpretation. In other words, this whole idea of the alternative being "too few subjects" is a matter for descriptive statistics, not inferential statistics.

In sum, there is no reason to test across subjects in a situation such as this, in which a one-tailed test is justified. Of course, the situation would be very different if we were doing a two-tailed test, for then we would be asking whether the population mean was or was not zero. Here the null hypothesis could be true if there were in fact opposing effects that cancelled out each other across the whole population of subjects, but happened to be weighted in one direction or the other in some particular sample of subjects chosen. Some other sample of subjects might yield a result in the opposite direction, a result that might even be significant across stimuli. But this situation is rare in experimental psychology, where most experiments are set up so that the hypothesis of interest produces an effect in one direction only, in spite of whatever mechanisms might be working in the opposite direction. (In other cases, one might do a two-tailed test when one would be satisfied with a one-tailed effect in either direction; for example, in the present case one might simply want to show that a phonemic code is available and has some influence on processing. In this situation, the present argument would apply, since this is not a "real" two-tailed test. Instead, the two-tailed test is being used in place of two post-hoc one-tailed tests. The critical point here is that in this case there is no real interest in the population mean for its own sake, but rather in whether *anybody* really deviates from it in either direction.)

Note that this is not an argument against *running* several subjects, only against *testing* across them once an effect has been shown across stimuli. A good reason to run a few different subjects is that some people will not engage themselves properly in the task and, thus, will show a certain kind of result for the wrong reasons. Thus, we usually throw out a subject whose reaction times are three times longer than everyone else's (of course, reporting that we have done so). Also, we may in fact be interested in descriptive statistics concerning the distribution of the effect over subjects.

An analogy may help drive home the point. Just as subjects, as a variable, can affect the magnitude of the result, so can times of the day. A difference between homophonic and nonhomophonic nonwords might be large (for all subjects and stimuli) at 2:00 p.m. but so small as to be nonexistent at 5:00 a.m. Yet we feel no need to sample various times of the day and test across times, subjects, *and* stimuli all at once.

This argument requires asymmetry between stimuli, on the one hand, and subjects or times of the day on the other, for it *is* still necessary to test across stimuli. Where is the difference? To find it, consider the extreme case in which we do the experiment with the entire population of subjects (all real and potential humans) and stimuli (all homophonic and nonhomophonic nonwords). Also, assume that the null hypothesis is true. Now, take any given pair of stimuli, one from each group, and look at the mean difference between its members averaged over all the subjects. Depending on the pair, it may be either positive or negative. Now look at the mean score for one subject across all stimuli. If the null hypothesis is true (and if the sample of stimuli is indeed infinite), this difference can only be zero or negative. If this difference is positive, then the mechanism which produces a positive effect must exist.

There are three practical implications of this argument. First, when a one-tailed test across stimuli is appropriate, it should be done in place of the tests Clark recommends, since it is both easier to do and more sensitive. Second, it is often a better research strategy to spend more effort sampling stimuli and less effort sampling subjects. And third, the situations in which one-tailed tests are appropriate may be more common than is often thought. Many experiments are best described as attempts to demonstrate the existence of mechanisms which produce an effect in a certain direction. When an experiment claims to ask which of two mechanisms is bigger or more powerful, there is often some argument that can be made to the effect that the conclusion would not apply for a different sample of stimuli, subjects, or presentation conditions. When such an argument can be made, all that can really be concluded is that the mechanism which produces the bigger effect exists.

REFERENCES

Clark, H. H. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 1973, **12**, 335-359.

Rubenstein, H., Lewis, S. S., & Rubenstein, M. Evidence for phonemic recoding in visual word recognition. *Journal of Verbal Learning and Verbal Behavior*, 1971, **10**, 645-657.

(Received for publication September 15, 1975.)