Statistical Science
1991, Vol. 6, No. 1, 100–116

The Philosophy of Multiple Comparisons

John W. Tukey

Abstract. This paper is based on the 1989 Miller Memorial Lecture at Stanford University. The topic was chosen because of Rupert Miller's long involvement and significant contributions to multiple comparison procedures and theory. Our emphasis will be on the major questions that have received relatively little attention—on what one wants multiple comparisons to do, on why one wants to do that, and on how one can communicate the results. Very little attention will be given to how the results can be calculated—after all, there are books about that (e.g., Miller, 1966, 1981; Hochberg and Tamhane, 1987).

Key words and phrases: Confident directions, confidence directions, multiplicity, priced-out data, split multiplicity, recombining value splitings, studentized birange double differences.

ESSENTIAL BACKGROUND

Significance

Statisticians classically asked the wrong question—and were willing to answer with a lie, one that was often a downright lie. They asked "Are the effects of A and B different?" and they were willing to answer "no."

All we know about the world teaches us that the effects of A and B are always different—in some decimal place—for any A and B. Thus asking "Are the effects different?" is foolish.

What we should be answering first is "Can we tell the direction in which the effects of A differ from the effects of B?" In other words, can we be confident about the direction from A to B? Is it "up," "down" or "uncertain"?

The third answer to this first question is that we are "uncertain about the direction"—it is not, and never should be, that we "accept the null hypothesis."

The follow-up question is about how much—about what we are confident of concerning the numerical difference

effect of A MINUS effect of B

which we shall abbreviate as A – B. If the first question was answered "direction uncertain," then the larger part of the answer to the follow-up question is how big might (A – B) be—what is the larger of the absolute values of the two ends of the confidence interval for A – B. The smaller part adds to this: "that's true in one direction, in the other (A – B) cannot be even that big, only so-and-so large!"

If the first question was answered "A – B positive," then the larger part of the answer to the follow-up question answers, usually: "What is the minimum size of A – B?". The smaller part, usually, answers: "What is the maximum size of A – B?" (Sometimes, but only rarely, these two are interchanged.)

If the first question was answered "A – B negative" since this is the same as "B – A positive," we have only to exchange the roles of A and B in what was just said.

We have to accept explicit uncertainty—initially about whether we are confident about direction, ultimately about the exact value of A – B.

Words, Thoughts and Actions

What of the analyst, who may even be a statistician, who says "This is all about words—I may use the bad words, but I do always think the proper thoughts, and always act in the proper way!"

We must reject such a claim as quite inadequate.

Unless we learn to keep what we say, what we think, and what we do all matching one another, and matching a reasonable picture of the world, we will never find our way safely through the thickets of multiple comparisons—and we will not serve ourselves, our friends, and our clients adequately.

Black and White—A Dangerous Dream

The worst, i.e., most dangerous, feature of "accepting the null hypothesis" is the giving up of

John W. Tukey is Senior Research Statistician, Princeton University, Fine Hall, Washington Road, Princeton, New Jersey 08544-1000.

explicit uncertainty: the attempt to paint with only the black of perfect equality and the white of demonstrated direction of inequality. Mathematics can sometimes be put in such black-and-white terms, but our knowledge or belief about the external world never can.

The black of "accept the null hypothesis" is far too black. It treats "between −101 and +1," "between −101 and +101," and "between −1 and +1" all alike, when their practical meanings are often very, very different.

The white of demonstrated direction of inequality is too white. On its face, it treats "between +1 and +101," "between +1 and +3," and "between +99 and +101" as if they were the same, when their practical meaning is quite different.

All too often, it is misused further when an observed difference of 50.1 is said to be highly significant, and this latter statement is, perhaps tacitly, interpreted as "believe all three digits of 50.1!" An observed value of 50.1 can be highly significant because we have tied the answer down between 20 and 80, or between 45 and 55, or between 49.8 and 50.4—tremendously different possibilities.

Black or white is dangerous, misleading, generally unsatisfactory. "Confident direction?" is the first question; something like a confidence interval —perhaps only the most relevant of that interval's two parts—is the badly needed answer to the follow-up question.

Knowledge or Belief, How Bought?

With what coin do we buy knowledge or belief? At least three different kinds of payment are always important: (1) The care and insight with which the data collection, or the experiments, were planned and performed. (2) The effort involved in collecting enough data. (3) The formal error rate that we are willing to accept for our conclusions.

The first two modes of payment are typically the responsibilities of our friends or clients, though we may be able to help with one or both. The third mode of payment needs to be a joint responsibility of investigator and analyst-statistician, who may be two people, or two roles for the same person. A clear understanding, both by the investigator and by the readers of the reports, of what has been spent is important—and a special responsibility of the analyst-statistician.

A 50% confidence interval will be shorter than a 95% or 99% confidence interval. If we are willing to spend a 50% chance of error, we can claim tighter knowledge than if we are willing to spend only a 5% or 1% chance of error. That extra "knowledge" —some would want to say, instead, that extra "belief"—was bought by accepting a greater error rate.

In the face of variable results (and when did we last have an instance when they were not variable) we can say little, if anything, without some chance of error. We must pay some chance of error to extract knowledge—or belief—from data. A crucial task is to expend this chance wisely, and to see that how it was spent, as well as how much was spent, is clearly recognizable.

Long ago, Fisher (1926, foot of page 504) recognized that truly solid knowledge did not come from analyzing a single experiment—even when that gave a confident direction with a very, very small error rate, like one in a million—but rather that solid knowledge came from a demonstrated ability to repeat experiments, each of which showed confident direction at a reasonable error rate, like 5%. This is unhappy for the investigator who would like to settle things once and for all, but consistent with the best accounts we have of the scientific method, which emphasize repetition, preferably under varied circumstances.

We do not dare work at very high error rates. We should not try to work at very low ones. We need to work in the range where error rates make an appreciable contribution to the "fuzz" that is always involved in our knowledge or belief. Just backing off from a single value to an interval, though very important, is not enough. We need to keep in mind the uncertainties that remain in the interval—some of them well measured by the corresponding error rate (or *diffidence*, where diffidence = 100% − confidence).

Our uncertainties are always possible: the data collection or experiment could have been misconceived, there may have been unrecognized incursion of other factors, and so on. Even if it were statistically sensible to analyze the data to give results with sharp edges, which it is not, doing this would be dangerously likely to distract our attention from such additional uncertainties.

Empirical knowledge is always fuzzy! And theoretical knowledge, like all the laws of physics, as of today's date, is always wrong—in detail, though possibly providing some very good approximations indeed.

An essential of what some call "the information revolution" is learning to live with—and made good use of—both fuzzy facts and theories that will be different later. As Kuhn (1970) has pointed out, such incomplete theories are a key tool in the conduct of science and its applications. The precise logic of mathematics serves statistician and data analyst in derivations—in theoretical structures

which do help us in thinking about the world. But how we think about the world needs to be suitably imprecise. We dare not limit ourselves to such formal precision.

Why Confidence Intervals?

We can all understand the importance of confident directions. Knowing which is more and which less—or which is higher and which lower—is clearly important. But why is interval knowledge important?

Four reasons seem to summarize the need for confidence intervals. In order of decreasing importance, they are: (1) the need for directions after pricing (or other adjustment); (2) the need to judge the compatibility of two or more studies of the same versions; (3) the need for rough indication of precision; and (4) the need to be able to adjust diffidence. A word about each is appropriate.

First, though we often tend to forget the fact, knowledge is gained, not only for its own sake, but for use. In use it has to be tempered.

It is not enough to know that 500 pounds per acre of a fertilizer will increase the yield of some crop. We want to go at least as far as "how much will it cost to buy and apply that amount of fertilizer?" and "how much will the extra yield return?" (We can hope that the value of the change in the condition of the soil remaining after the crop is lifted is also considered!)

For a given *pricing scheme*—a given set of prices for fertilizer, application, harvesting and storage, sale of crop—there will be a break-even increase in yield (for the given fertilizer application). What we would like to be confident about is how the increase in yield relates to this break-even point, which surely will vary from year to year, as prices and wages vary.

Indeed, for whatever year is next, we will not know crop prices in advance, and we may need to ask about confident directions as seen from two, three, or more possible break-even points, corresponding to two, three, or even more pricing schemes.

A confidence interval offers us answers about confident direction assessed from any, or all, break-even points. Nothing simpler can do this.

Many of us are familiar with *deriving* a confidence interval from an infinite array of tests of significance, one for each potential null hypothesis. Fewer of us, perhaps, have thought of the *use* of a confidence interval as the reverse process. This is the most important reason for a confidence interval, and developing such an interval demands as much care as procedures to study confident direction can provide.

Second, though possibly more important than (1), is the need to compare the results of two (or more) studies involving (some of) the same versions, asking whether the results are compatible. Knowing that a comparison had the same confident direction in two studies does not ensure that the results are compatible (as far as that comparison goes); they may be compatible, or they may not. Knowing that a comparison was of uncertain direction in one study, but of confident direction in another, does not ensure that the two studies are incompatible; they may be compatible, or they may not. Confident direction is not enough. We need confidence intervals.

Third, as we have already implied by contrasting (1, 101) knowledge with (99, 101) knowledge, we do need to know at least roughly what kind of precision is involved in our knowledge or belief. It is often hard to separate, in our minds, this need from the first need. We tend to think of standard errors, for instance, as being there to construct intervals. But, if we try hard, we can distinguish the third use, and find that, for that use, rather rough values will ordinarily meet our needs. (This is fortunate, for we rarely know precision other than roughly.)

Some would say this third use is more important than the second. For those concerned with techniques of statistics and data analysis, this is probably so. For those concerned with the real world, it seems doubtful.

Fourth, there is a more technical reason. If the 95% confidence interval goes from 70 to 130, a 99% interval will be wider, but probably not much outside 60 to 140, whereas a 50% interval will be narrower, going roughly from 90 to 110. At least roughly, a confidence interval for one error rate tells us about confidence intervals for other error rates.

Nothing like this works for confident direction. If we are confident that $A > B$ at 5% error rate, we have no idea whether the same inequality holds for a 1% error rate or not. If we are uncertain of the direction of $A - B$ at 5% error rate, we have no idea whether we would be confident of the direction at 50% error rate or not.

It should now be clear (a) that confidence intervals are irreplaceable and (b) why they are needed.

Confident Conclusions versus Interesting Hints

We have stressed the fuzzy character of our knowledge and the roles of confident direction and confidence intervals. We have emphasized the variety of "direction uncertain": how $(-101, 1)$, $(-1, 1)$, and $(-100, 101)$ differ greatly, even if they are all for the same error rate. Clearly, then, it would be a mistake to say nothing more about all the compar-

isons for which, at our chosen error rate, we have to say "direction uncertain."

Some deserve nothing more, others need "although imprecisely measured" or "even though quite precisely measured," and still others deserve recognition as hints. Thus $(-1, 101)$, for instance, is likely to deserve something like "direction uncertain, but plausibly positive."

We have not thought enough about hints, either informally or formally. Catherine Marsh and I are trying, from time to time, to make progress here. Never taking any notice at all of hints, however, is clearly very nearly the worst thing we could do.

SIMPLE MULTIPLE COMPARISONS

The Challenge

A man or woman who sits and deals out a deck of cards repeatedly will eventually get a very unusual set of hands. A report of unusualness would be taken quite differently if we knew it was the only deal ever made, or one of a thousand deals, or one of a million deals, etc.

Someone who raises 1415 strains of some variety of corn and measures the yield or each can make slightly more than a million comparisons of the yield of one strain with that another. Surely the largest of these will tend to be larger than the largest comparison found when only 46 strains were measured, so that only a few more than one thousand comparisons could be made.

We clearly need to think, and think hard, about how to handle such questions of multiplicity. And we cannot expect unique answers. For what we think has to depend on how the results are accessed or analyzed.

At one extreme, the 1415 results may be just stored away, available, like birth certificates, for a small fee. If only an occasional pair of gamblers pays the fee, so they can settle a bet on which of two strains yielded higher, the essential multiplicity will be somewhere between "one" (if we think of one pair of purchasers at a time) and the number of purchasing pairs—surely far less than a million.

At the other extreme, where an analyst puts the 1415 results in his or her PC and uses lots of data analysis software, we can be sure that the most unusual of the million comparisons will be found and, if at all unusual, reported upon. Exploration—which some might term "data dredging"—is quite different from "exogenous selection of a few comparisons." Both have their place. We need to be prepared to deal with either.

One of the easy ways to deal with either is to be careful about the definition of error rate. As a rate, an error rate has a numerator (a number of "errors") and a denominator (a number of "trials"). For one exogenously selected comparison:

(I) "error" = getting that particular comparison wrong
"trial" = repeating data collection for those two treatments,

while for exploration of the million comparisons, followed by emphasis on the largest of all (done properly, good science—in fact essential to good science), either

(F) "error" = getting one or more of the million comparisons wrong
"trial" = repetition of data collection for all 1415 "treatments"

or

(B) "error" = count one for each comparison wrong
"trial" = repetition of data collection for all 1415 treatments.

If we apply a conventional procedure, with a single-application error rate of 5%, we will obtain different error rates for the 3 definitions, namely

(I) error rate = 5%
(F) error rate $\geq$ 99.9%
(B) error rate = 5,000,000%
(an average of 50,000 errors/trial).

If either of (F) or (B) is appropriate, we must do something. We could go in either of two ways:

- reduce the individual (I) error rate to make the family-wise (F) error rate or the Bonferroni (B) —or per family—error rate tolerable,
- compare the total number of unusual comparisons found with the 50,000 expected by pure chance.

In the present—very extreme—case, we may rightly wince at working to extreme at an individual error rate as 0.000005%. Is there any hope that our conventional statistical procedures will provide useful confidence intervals (or tests of significance) corresponding to such extreme error rates? Who knows? But, if we were to use simple robust statistical procedures, such as the biweight-based analog of Student's t, the results of Kafadar (1982) indicate that we could do all right down to error rates at least as small as 0.001%.

One-stage experimentation or data collection with fourteen-hundred-odd candidate treatments may well be silly; there may be no substitute for picking

out the 50 or 100 apparently most extreme candidates and repeating an independent data collection for this smaller collection. But for more reasonable numbers of candidate treatments, we can do quite well with making the individual error rates appropriately small.

The other approach, sometimes called "the higher criticism," is insufficiently rewarding. Suppose we find ourselves individually confident about the directions of 60,000 of the comparisons. Pure chance would give 50,000 ± (something like 218) of these. Clearly there is a significant excess; indeed, rather more than 15% of the 60,000 observed did not arise by chance. But which 15%? Individual knowledge, which is what we are likely to need, is prominent by its absence.

Thus, in the situation considered, we need to follow the first approach, making more conservative statements about each of the comparisons.

So far we would appear to have acted as if the different comparisons were uncorrelated, even stochastically independent. For the first approach, this is only an appearance. Error rates arise by adding up things, and the average value of a sum is the sum of the average values, no matter how things are correlated.

When we allow for correlation in the second approach, the correct number in "± something like 218" will differ from 218 by a small or moderate factor. Once we calculate this factor, it is trivial to revise the example.

Less Extreme Cases

Comparisons build multiplicity fast. Even though we do not expect the million comparisons of our somewhat fanciful example, it is easy to get enough comparisons to matter, as Table 1 shows.

Directions or Intervals

Again, we must return to the choice between directions and intervals. Which do we need?

For pure intellectual curiousity—and perhaps for writing treatises for the intellectually curious—it may be that confident directions (up, uncertain, or down) can suffice. It would be good to understand why many psychologists, for example, seem to be content with only confident directions. Is this a desire for abstract knowledge? Or a sign of inability to make use of more quantitative results? Or an unwillingness to price (or a lack of experience in pricing) comparisons as a basis for real-world actions? Or a belief that *qual*itative knowledge is all that psychologists can hope for? Or what?

We need to understand the lure of confident-direction multiple comparisons; and we must be prepared, in the interim, to see an industry continuing to develop such procedures (an industry to which I have contributed).

TABLE 1
Multiplicity for simple comparisons

Number of candidates	Number of comparisons	Corresponding change[a]
4	6	5% → 1%
5	10	5% → 0.5%
7 or 8	21 or 28	5% → 0.2%
10 or 11	45 or 55	5% → 0.1%
15	105	5% → 0.05%
23	253	5% → 0.02%
32	496	5% → 0.01%
45	990	5% → 0.005%

[a]From simultaneous error rate to approximate corresponding individual error rate.

For the four abilities discussed above—direction after pricing, compatibility of parallel studies, rough indication of precision, adjustment of error rates—confidence-interval multiple comparisons seem essential, if the results of the analysis are intended to guide practical actions.

Split-Multiplicity

In many situations we could look at any or all of many things, and would if the amount of available data were not severely restricted. One of these is clinical trials—of drugs or therapies—where both ethical and financial considerations limit the size of the trials. In such situations, it is often essential to focus sharply on only one or two primary questions, questions that deserve analysis in terms of confident directions (and which often may as well be analyzed in terms of confidence intervals). We can then spend all (if there is one question) or half (if there are two) of our permissible error rate, often 5% overall, on the single primary question, or on the two primary questions.

Once we have spent this error rate, it is gone. And what we say about the remaining questions has to have many of the properties of *hints*, even if we work at, say, an individual error rate of 5%. Keeping the different strength with which we believe primary and auxiliary answers—especially when these answers appear to use the same statistics (e.g., Student's t) in the same way—is a very serious and important challenge.

The message has to be that it can be wise and necessary to focus on a very few *prespecified* questions, prespecified before data collection, whenever we cannot enjoy the luxury of enough data to work

with either familywise (F) or Bonferroni (B) error rates.

The Extremes

We can thus specify a small number of relatively extreme situations for which we need to be prepared.

(1) *The data bank, exogenously used*, where only a few of the many values will ever be looked at, and the relevant multiplicity is the few that are looked at.

(2) *The clinical trial, focused, but not exclusively*, where a very few prespecified comparisons will be allowed to eat up the available error rate, and the remaining comparisons have the logical status of hints, no matter what statistical techniques may be used to study them.

(3) *The full exploration*, where we may, for instance, look at all simple comparisons, shrinking our per comparison error rate enough to keep our simultaneous error rate, whether (F) or (B), at the desired value.

The Studentized Range

If we desire to be ready to assess direction after any possible pricing scheme has been applied, then exchangeability is a reasonable approach. This is so because, looking at all possible pricings, for whatever set of $z_1, z_2, \ldots, z_n$ you have in mind, perhaps as a standard, there will be prices c_i such that the observed y_i, referred to the prices, give

$$\begin{aligned} z_1 &= y_1 - c_1 \\ z_2 &= y_2 - c_2 \\ &\cdots \\ z_n &= y_n - c_n \end{aligned}$$

and the same will be true for any permutation of the z's. Thus, for all possible uses of direction after pricing, one pattern of candidate values is like any other pattern. Until we know the prices, we do not know which pairs of candidates are close together, and which are far apart.

If $\eta_1, \eta_2, \ldots, \eta_n$ are the long-run values that $y_1, y_2, \ldots, y_n$ are estimating, the largest "error" of a comparison is

$$\begin{aligned} &\max_{i,j} \left| (y_i - \eta_i) - (y_j - \eta_j) \right| \\ &\quad = \max_i (y_i - \eta_i) - \min_j (y_j - \eta_j) \\ &\quad = \text{range of the } (y_j - \eta_j). \end{aligned}$$

If we can have a bound for this range, say of the form range $\le qs$, where q is a tabulated critical value of the Studentized range and s is an appropriate estimate of error, we can express this bound range $\le qs$ in the form

$$(y_i - y_j) - qs \le \eta_i - \eta_j \le (y_i - y_j) + qs,$$

which holds simultaneously for every pair (i, j), for every one of the $n(n-1)/2$ comparisons.

Thus what is probably *the* most natural way of setting confidence intervals for all (simple) comparisons leads to the use of the Studentized range.

At the other extreme, we would like to ask whether we believe that any comparison has been shown to differ from zero. This means, probably, asking whether the comparison that seems the largest—the difference between the lowest observed y and the highest observed y—has been shown not to be zero. (This is essentially a question about direction with all prices at zero!)

Again, the simplest answer, and probably the best one, uses the Studentized range. Thus the innocent may be pardoned for thinking that, if the two extreme situations shout "Studentized range, huzzah!" this approach ought to serve as well for all intermediate cases. But this does not seem to be the case.

Just why this is so is not yet crystal clear to me. It may well be that "any and all pricing systems" for which the confidence intervals work is so different from "all prices zero" that almost nothing lies in between.

As the discussion of the close of the paper indicates, we should think of the Studentized range as an example of a technique, rather than as something inevitable.

The Confident-Directions Industry

We have admitted that we must anticipate such an industry (a cottage industry?) to continue producing new confident-directions procedures. We need to ask: How are the various confident-directions products differentiated? Why is there a place for more than one?

Essentially the fit into different stages in the fineness of resolution of measurement in a field. At an early stage, we are happy to find even one pair of treatments whose results differ in a definite direction! Later, we expect to find several pairs, and we want to increase, at least somewhat, the numbers that we find. Eventually, we want to find *all* the pairs different and want to focus our effort on separating the most difficult pairs. For different stages along this path we can use different confident-directions procedures.

Thus it seems reasonable to use: (1) the studentized range (or a Studentized maximum absolute

deviate) near the beginning: (2) Roy Welsch's (1977) gap-and-stretch techniques toward the middle; and (3) such procedures as those of Peritz (1970), Ramsay (1981), and even the more recent addition by Braun and Tukey (1983), at a quite late stage.

Other Contrasts

Very early in multiple-comparisons history, Scheffé (1953) pointed out that using F instead of the Studentized range would have confidence intervals for every contrast (for every $\sum c_i y_i$ with $\sum c_i = 0$) with the property that

- for every contrast the simultaneous interval as the same multiple of the individual interval.

A decorative property, which some found morally appropriate or even ethically required.

However, (a) doing this widened the confidence intervals for simple comparisons; (b) it was often hard to find contrasts—other than comparisons—in which there was much interest; (c) most of those found were comparisons of subgroup means, for which the s^2 for simple comparisons was usually NOT an appropriate error term; and (d) limits on all simple comparisons, when the Studentized range was used, rather than F, implied, indirectly, limits on all contrasts (weaker for general contrasts than those coming from F).

As a result, the F-approach has been quietly laid to rest, leaving other contrasts to the weaker indirect consequences of the Studentized range. Today, we might let the pendulum swing back very slightly, and qualify the relegation to (d) by "unless a specific contrast—or each of very few specific contrasts—is either prespecified or exogenously identified, when it can be treated on a par with the simple comparisons."

GRAPHICS FOR THE SIMPLE CASE

Determinations: Single Arrows and Extensions

It seems natural to show a single confidence interval with either a double-ended arrow or an aperture, either by the presence of a short black mark or by the absence of a middle section of a long black mark. If we want to show 2 distinct confidence intervals, ordinarily one inside the other, we can combine the double-ended arrow for the shorter interval with extensions of some sort to delineate the longer interval.

The simplest settings in which to illustrate this involve *multiple determinations*, where each result is compared with an external or fitted reference, rather than multiple comparisons, where results

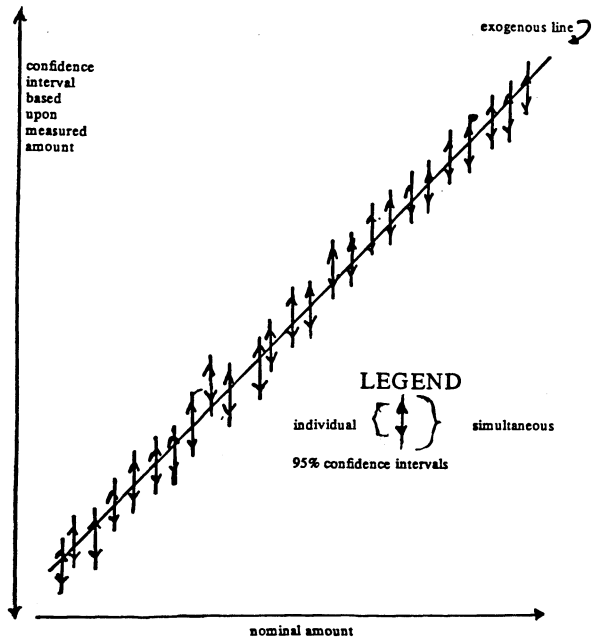

FIG. 1. *Hypothetical calibration data (individual and simultaneous confidence intervals).*

are compared in pairs. Figure 1 shows a hypothetical example, where 25 confidence intervals (individual and simultaneous) are compared with a fitted line.

In this example 24 of the 25 *individual* confidence intervals, each shown by a pair of arrowheads and the vertical line segment joining them, cover the line (if they were 95% intervals and the line were perfect, we would expect an average of 23.75 coverages—95% of 25). All the *simultaneous* intervals, each shown by extended line segments, cover the line (if they were 95% simultaneous intervals, this ought to happen an average of 19 times in 20 complete repetitions, each with 25 simultaneous confidence intervals).

If we work with intervals shown by apertures, as openings between long bars (or segments of bars), we can easily show more kinds of confidence intervals in one picture.

Using apertures is a particularly truthful approach, because the emphasis on bars (or termini of bars) stresses the potential values that have been ruled out, those that we know most about, because we are ruling them out! (Most of the values in the confidence interval are incorrect, but we preserve them all, because we do not know which one is correct! Thus the in-interval values are always an unknown mixture of "yes" and "no.")

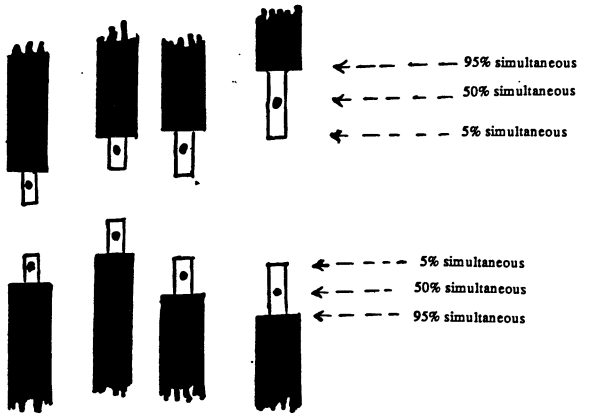

FIG. 2. *Simultaneous (confidence) aperture plots at three levels (alternative diaglypt).*

Multilevel Simultaneous

If we take simultaneous confidence intervals very seriously, we will find use for more than one choice of simultaneous error rate. Having 5% error rate, as in 95% simultaneous intervals, need not be enough.

We would also like to know when the fit is too good, as when all 5% simultaneous intervals (with a 95% simultaneous error rate) cover the reference.

And there can be interest in whether the fit is "worse than average" or "better than average," so that 50% simultaneous intervals can be helpful.

Figure 2 shows what a piece of such a 95%/50%/5% simultaneous display might look like.

Comparisons: Notches; Underlaps versus Overlaps

When we come to comparisons, we need to pause and take stock. If we are to compare n candidates, there will be $n(n-1)/2 = \binom{n}{2}$ comparisons. No one wants to look at a picture with that many things in it. How can we boil things down better?

If the intervals have a common length, we can assign half this length to each candidate, in such a way that the whole length is formed by combining each of the two candidates' halves. This calls for attaching $\pm\sqrt{2}$ SE (where SE is a standard error for individual candidates' results) to the observed value for each candidate and thinking of differences corresponding to all four combinations of $\pm$ for one with $\pm$ for the other. The two extreme of the four differences, gotten by combining one + with one −, will then provide the ends of a confidence interval, located at $\pm 2\sqrt{2}\,\mathrm{SE} = \pm 2$ (SE of difference) from the observed difference. This will

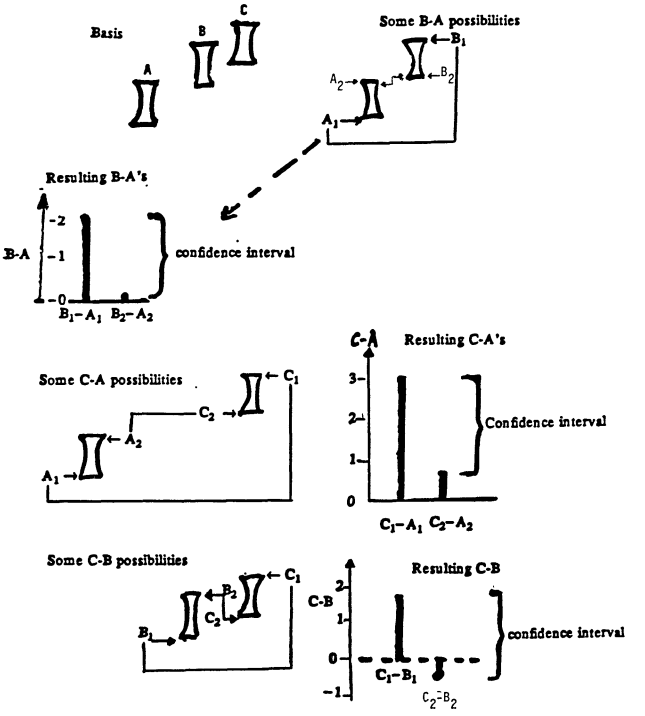

FIG. 3. *The fundamentals of the notch approach.*

be close to a 95% interval, and we can adjust the factor of 2 to whatever we need.

In Figure 3, the $\pm\sqrt{2}\,\mathrm{SE}$ intervals are shown as smooth hollow notches. The exhibit illustrates the construction of the three confidence intervals.

If our concern is only for confident direction, then everything is quite simple. (1) If two notches overlap, then we are *not* confident about the direction from the value for the first candidate to the value for the second. (2) If the two notches do *not* overlap —if they *underlap*—we *are* confident about the direction from one candidate's value to the other.

Supplementary Comparison with a Scale

It is easy to verify that the notches we want for comparisons are only 0.7071 as long as the intervals we would want for determinations. (The looseness of a comparison corresponds to twice such a length, hence to 1.4142 times that for a determination, as it should.)

If we want to also use our notches as determinations, comparing individual observed results with numerical values, then we need to give those numerical values some looseness of its own. How much? Clearly

$$\pm 2(1.0000 - .7071)\mathrm{SE} = \pm .586\mathrm{SE}.$$

(We leave designing the picture as a challenge.)

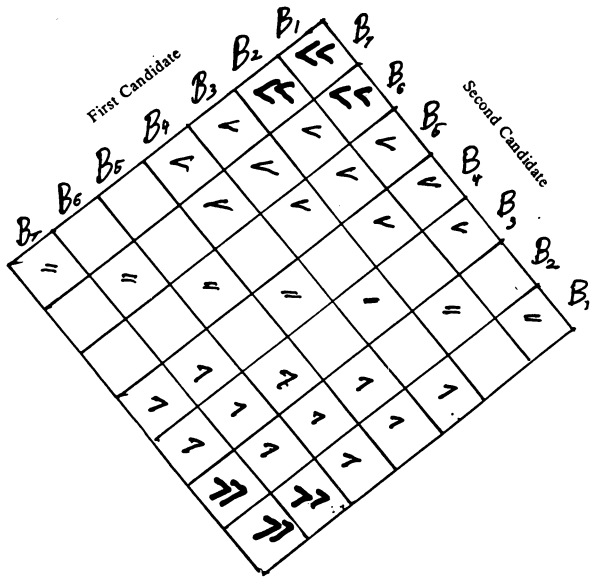

FIG. 4. *Semigraphical display of directional information.*

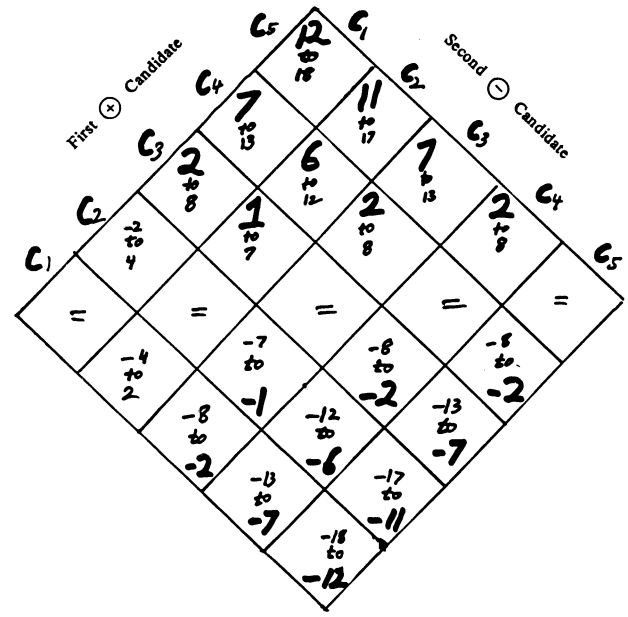

FIG. 5. *Semigraphical display of confidence intervals for differences (comparisons) stressing inner ends.*

Semigraphical n^2 Displays (for not too large n)

Suppose, again, that we have n candidates to compare. Can we show the same sort of information more explicitly? What would be a good format? Figure 4 shows a rotated two-way table, one candidate versus the other, for seven candidates, where each cell contains

(a) " > > " if we are strongly confident that first > second,
(b) " > " if we are confident that first > second,
(c) " " if we are not confident about direction,
(d) " = " if the two candidates are the same,
(e) " < " if we are confident that first < second, and
(f) " < < " if we are strongly confident that first < second.

Clearly such a picture does quite well in expressing directional information.

Figure 5 uses another form of the 45° approach to show confidence intervals for all differences of pairs of 5 candidates, using size and boldness of numbers to (a) stress the inner ends of the interval and (b) give a "blackness only" display of directional information.

More Complex Schemes

More complex schemes, some easy to improve, are in use in some subject-matter areas, but discussing them would take us too far from our main threads: (a) that simple graphic displays of multiple comparison results are not too hard to find and use: and (b) that looking frequently at such displays is likely to improve our insight into—and the general usefulness of—multiple-comparison procedures.

ANALYSIS OF VARIANCE, POSSIBLY COMPLEX

Alternative Purposes

Analysis of variance (we will confine ourselves here to data in a balanced factorial pattern) began as an implicit breakdown of the data, as reflected explicitly by only mean squares, one for each "source of variation" considered. Computation was expensive, and incomplete breakdowns (e.g., leaving a variety of higher-order interactions in the "error") were frequent.

Today computation is cheap, and a complete breakdown—into common, main effects, two-factor interactions, three-factor interactions, and so on—is easily affordable. To each possible mean square, then, will correspond a table of values. If we make enough copies of such a table to cover, when placed side-by-side, the initial data pattern, we obtain what is conveniently called an *overlay*. If we think of first laying these overlays on top of one another, and then adding up the entries that lie above one another, we regain the original data. Thus the overlays define a *value-splitting* of the original data. Ease of computation has made it feasible to enhance the mean squares with detailed breakdowns that are likely to be much more informative. Even 25 years ago, of course, it was common to criticize papers that claimed significance for the main ef-

fects of some factor without quoting the means for the different versions of the factor in question. (A quantitative-valued factor has *levels*, but it is misleading to use "levels" for such things as sex, variety, or operator. So we use *version* as the general term for the "value" of a factor.)

Initially, one breakdown (often incomplete) was made implicitly, and summarized by mean squares. Complete breakdowns are now easy, exposited by tables and still summarized by mean squares. A modern approach, which we shall mention only generally, involves recombining overlays where, for instance, a main-effect table seems to consist mainly of summarized interactions. Such combinations would be regarded as forced when the analysis is essentially exploratory, and only the recombined tables would be discussed.

Recombination also arises at the opposite extreme. If, for instance, differences of main effects for some factor are clear as to direction, our next step could be to ask whether this is also the case conditional on some version of another factor. The relevant table of values induces recombining, for a further analysis only, the main effects of the first factor with the interactions of the two factors.

Structure

Among uses the analysis of variance, different applications vary widely, some extremes being:

- Prespecified comparisons, which call for more or less complete splitting of each data value into parts, one part for each line in the prescribed analysis (the resulting subtables are conveniently called *overlays*).
- Full exploration, which calls for, first, a very complete value-splitting, followed by a selective recombination (guided, not by apparent significance, but by whether a sufficiently large fraction of the overlay in question seems to be noise).

Whatever approach may be appropriate in a particular situation, the character of the (intermediate) results is the same:

- Some one-way tables, and a standard error appropriate for comparisons.
- Usually, one or more two-way tables, and standard errors appropriate for comparisons within them (within rows, within columns, possibly crosswise).
- Often, one or more three-way tables, with appropriate standard errors.
- Sometimes, one or more four- or more-way tables, etc.

The one-way case is simple multiple comparisons, as discussed above. The two-way case will be discussed next, leaving the three- and more-way cases for other times and places.

Each one-way table started as a table of summaries, each entry summarizing all the values where a particular factor appears in a particular version. (Summarization classically by arithmetic means.) Subtraction of the common term then makes main summaries into main effects, which we study in terms of main comparisons differences of pairs of main effects, which we are also differences of main summaries.

Each two-way table also started as a table of summaries, each entry summarizing all the values where a particular pair of factors appear in a particular pair of versions. (Summarization also classically by arithmetic means.) Subtraction of both main effects, and of the common term, converts two-way summarizes into two-way interactions. For a specific purpose, we may need to analyze the table of two-way summaries, or perchance the table of two-way interactions, or some intermediate table, say one conditional on one of the factors.

How to Describe What We Calculate

The description of alternative extremes just given will not seem familiar, even to those who feel accustomed to work with the analysis of variance. One reason for this goes back to the early decades of analysis of variance, when computing was effortful and expensive; when saving arithmetic was vital, when we could just barely afford to calculate mean squares. What could one do with mean squares? Form F-ratios and announce the results of significance tests seemed the only answers, but somehow people made better use of analysis of variance than these answers—and the formalizations they spawned—could support.

Today we are still actively learning about how to describe analysis of variance as a flexible tool for understanding data. Although the process may still be far from complete, we have progressed far enough to make a very different description worthwhile.

So let us consider a two-way table $\{y_{ij}\}$ of responses, and think about a decomposition

$$\{y_{ij}\} = \{c_{ij}\} + \{a_{ij}\} + \{b_{ij}\} + \{d_{ij}\},$$

where $c_{ij} = c$, $a_{ij} = a_i$, and $b_{ij} = b_j$, which splits up the numerical value of each y_{ij} into four parts according to

$$y_{ij} = c_{ij} + a_{ij} + b_{ij} + d_{ij},$$

The raw table

	B_1	B_2	B_3	B_4	B_5
A_1	19	5	11	14	11
A_2	5	2	4	3	1
A_3	-12	-7	-9	-8	-9

data

The row and column summaries (here means)

					12
					3
					-9
4	0	2	3	1	

A* main summaries

B* main summaries

The main effects

					10
					1
					-11
2	-2	0	1	-1	2

A main effects

B main effects

common

The classical splitting

5	-5	-1	1	0	10
0	1	1	-1	-1	1
-5	4	0	0	1	-11
2	-2	0	1	-1	2

A × B interaction

A main effects

B main effects

common

Common, A and A × B recombined

17	7	11	13	12
3	4	4	2	2
-14	-5	-9	-9	-8
2	-2	0	1	-1

A @ B conditional effects (by column)

B main effects

FIG. 6. *An illustrative two-way table, part* 1. (*Numbers to the left, identification to the right.*) *Note that recombination was not forced on us and that we are not forgetting about the main effect of A. Rather we are recombining so that we may supplement our attention already given to that main effect with attention to the corresponding conditional effects.*

which can also be written

$$y_{ij} = c + a_i + b_j + d_{ij},$$

a value-splitting often made unique by requiring

$$\sum_i a_i = 0 \qquad \sum_j b_j = 0$$

$$\sum_i d_{ij} = 0 \qquad \text{for all } j$$

$$\sum_j d_{ij} = 0 \qquad \text{for all } i.$$

The conventional language now calls $\{a_i\}$ and $\{b_j\}$ sets of *main effects* (for factors A and B, respectively) and $\{d_{ij}\}$ a set of interactions (for the interaction A × B). We shall go further, not only thinking of the jth column of d_{ij} as A × B_j—as the supplement to the main effects of A that applies to version j of factor B—but also focusing on

$$\{c + a_i + d_{ij}\} = \{c_{ij} + a_{ij} + d_{ij}\}$$

as made up of the conditional main effects of A. The jth column of this last two-way table, whose entries are

$$c + a_i + d_{ij}$$

for some fixed j, tell us about the behavior of the response as a function of i, *conditional* upon factor B being in version j. It is natural to label this column as A @ B_j. It follows that

$$\text{A @ } B_j = \text{A} \times B_j + \text{A} + \text{common} = \text{A} \times B_j + \text{A*},$$

where A* = A + common is a vector of main summaries.

Before turning to the multiple-comparisons questions which this sort of structure generates, a numerical example can help to fix the ideas.

Two-Way Tables Illustrated

Figure 6 shows a 3 × 5 table of responses, the two one-way tables and one two-way table into which this 3 × 5 table is classically split, and the result of recombining one of the one-way tables with the two-way table. (We have chosen to recombine A as an illustration of how conditional comparisons relate to unconditional ones. In the sort of situation illustrated we would *begin* by looking at the main effects of A, not by eliminating them by recombination.) Figure 7 shows each overlay as a two-way array, and how they add up. Figure 8 continues this example with a variety of vertical-column-of-3 tables and subtables, and two versions of the same double difference or bicomparison (explanations below).

In Figure 8 the vertical-3's have been equipped with standard errors, the one for A being calculated from the A × B interaction and those for the subtables of the two-way tables being supposed—on the basis of other evidence not shown—to be small.

One of the first questions about such a data set will be: For which main comparisons—which comparisons among the main effects of A—are we confident of direction? (In our illustrative example, where all such differences are more than four times their standard error, we will be confident of the direction of all three main comparisons among main effects of A.)

Once we have dealt with the one-way tables (= the sets of main effects), it is time to look at the two-way table. We shall begin by looking at columns. Four columns-of-3 are relevant to the first column of the A × B interaction. They are interrelated, as shown in exhibit 8, by

A @ B_1		common		A		A × B_1
17	=	2	+	10	+	5
3		2		1		0
−14		2		−11		−5

We have already dealt with the main effect of A, and we usually don't look at the common. We ought to ask, "Which of the other two deserves our attention next?"

We can answer this on grounds of simplicity and purity. The left-hand column above—the effects of A at B_1, which we may write A @ B_1— depends only upon what was observed at B_1. However, whenever there is a real A × B interaction, the main effects of A will depend upon which other versions of B, beyond B_1, are present. Accordingly, the supplementary effects of A at B_1 will depend upon what was observed at B_2, at B_3, ... Clarity and simplicity thus favor looking first and foremost at "the conditional effects of A at B_1"!

Moreover, if we ask which kind of column is of the greater practical importance, we are led to the same answer. It can be highly important to know in direction (or in both direction and amount) what changing A will do when B = B_1.

Real data examples show that it can be interesting to know about direction, etc., for the additional effect attributable to A when B = B_1, as compared with when averaged over all 5 versions of B. Sometimes, it seems reasonable to expect, these directions may prove important, and not just interesting.

After we have looked at all the conditional effects of A—the effects of A at B_1, at B_2, . . . , and at B_5—we may want to go further, and ask about

The full overlays, splitting each of 15 y_{ij} into parts

19	5	11	14	11	$\{y_{ij}\}$
5	2	4	3	1	
−12	−7	−9	−8	−9	
=					=
2	2	2	2	2	$\{c_{ij}\}$
2	2	2	2	2	
2	2	2	2	2	
+					+
10	10	10	10	10	$\{a_{ij}\}$
1	1	1	1	1	
−11	−11	−11	−11	−11	
+					+
2	−2	0	1	−1	$\{b_{ij}\}$
2	−2	0	1	−1	
2	−2	0	1	−1	
+					+
5	−5	−1	1	0	$\{d_{ij}\}$
0	1	1	−1	−1	
−5	4	0	0	1	

FIG. 7. *An illustrative two-way table, part 2.*

comparisons within the columns of A × B—but this will be a secondary set of comparisons.

Preferred Questions about Two-Way Tables

Far too little attention has been given to the nature of appropriate and detailed analysis of two-way tables. The same sort of care (and emphasis on interpretability) that leads to split-multiplicity in clinical trials and comparison-focused analysis of one-way tables leads here to focusing on two or three kinds of comparisons and to allowing many simple comparisons to fall to second- or third-class status—unless rescued by prespecification or exogenous identification.

Knowing the direction from "version 2 of factor A with version 4 of factor B" to "version 3 of factor A with version 1 of factor B" is clearly not easily to interpret. Such a comparison does not deserve first-class, or even second-class status, unless attention has been called to it in some exogenous way that also provides an interpretation for results about it.

In studying two-way tables, first-class status goes automatically to:

- comparisons within a column (or within a row), naturally called *conditional* comparisons; their interpretations parallel those of the corresponding main comparisons, with averaging over the second factor (and other factors) replaced by conditioning on a version of the second factor (and still averaging over the others), and
- *crossed* comparisons (*bicomparisons*), which illustrate interactions with the same sort of specificity and relative simplicity that main comparisons provide for main effects (crossed comparisons are the same for tables of summaries as for tables of interactions, or for anything between),

with second-class status for:

- comparisons within a column (or within a row) of the 2-way table of interactions—these are naturally called supplementary comparisons, and their interpretation is ordinarily as the changes from main comparisons to the corresponding conditional comparisons (accordingly they are usually of less interest).

More general simple comparisons—and all other contrasts except crossed comparisons—deserve at best third-class status. Since any comparison, and hence any contrast, among summaries can be written as a linear combination of conditional comparisons, any set of confidence intervals for all conditional comparisons impllies, as logical consequences, confidence intervals for comparisons and all contrasts. These intervals give rather weak answers, but what more do third-class questions deserve?

Error Rate Issues

We could easily ask for any one of quite a number of error rate behaviors, to apply to our comparisons within conditional effects of A. We need to think carefully and experiment broadly with different choices. Here, I shall illustrate just one choice: 5% simultaneous for all conditional effects of A combined. With 5 columns (A @ B_1, A @ $B_2, \ldots,$ A @ B_5) we can afford (à la Bonferroni) an error rate of 1% for each column. The easy way to spend this is to use, in each column separately, the 1% point of the Studentized range of 3.

If we want to also include some attention to the columns of A × B, we have even more choices. One (of many) that maintains 5% simultaneous over all 5 + 5 columns-of-3 would use the 3/4% point of the Studentized range of 3 for each column of A @ B (thus spending 3/4 of 5% on the conditional effects of A) and the 1/4% point of the Studentized range of 3 for each of the columns of A × B (thus spending the remaining 1/4 of 5% on the supplementary

Some vertical columns of 3: (standard errors of individual entries at foot)

common	A	A×B$_1$	A×B$_2$	...	A @ B$_1$	A @ B$_2$
2	10	5	−5	...	17	5
2	1	0	1	...	3	2
2	−11	−5	4	...	−14	−7
	± 1.72	± small	± small		± small	± small

Some relationships among columns of 3:

common		A		A×B$_1$		A @ B$_1$	...	common		A		A×B$_2$		A @ B$_2$
2	+	10	+	5	=	17	...	2	+	10	+	−5	=	7
2	+	1	+	0	=	3	...	2	+	1	+	1	=	4
2	+	−11	+	−5	=	−14	...	2	+	−11	+	4	=	−5

One crossed comparison = one double difference = one bicomparison:

5	**−5**	−1	1	0	(5)−(−5)−(−5)+(4) = 19	(based on interaction)
0	1	1	−1	−1		
−5	**4**	0	0	1		

19	**5**	11	14	11	(19)−(5)−(−12)+(−7) = 19, also!	(based on data)
5	2	4	3	1		
−12	**−7**	−9	−8	−9		

A = main effects of A
A×B$_1$ = interaction of A and B at B$_1$ = supplementary effects of A at B$_1$
(supplementary to the main effects of A)
A @ B$_1$ = effects of A at B$_1$ = (B$_1$) conditional effects of A

FIG. 8. *An illustrative two-way table, part* 3.

effects of A). (Often other splits than 3/4, 1/4 would be reasonable.)

Crossed Comparisons, Double Differences

The elementary form of an *interaction* is a crossed comparison (a double difference, a bicomparison) that is not zero. This may be either an appearance that

$$y_{ij} - y_{ib} - y_{aj} + y_{ab} \neq 0$$

or an underlying fact that

$$\eta_{ij} - \eta_{ib} - \eta_{aj} + \eta_{ab} \neq 0,$$

where we must have

$$i \neq a \quad \text{and} \quad j \neq b$$

(else the double differences vanish identically).

The interaction character of these $\neq 0$'s is clear from their two equivalent forms: first

$$y_{ij} - y_{ib} \neq y_{aj} - y_{ab}$$

or

$$\eta_{ij} - \eta_{ib} \neq \eta_{aj} - \eta_{ab},$$

which says that the net effect of changing the 2nd factor from j to b is different when the 1st factor is at i than when it is at a; and second

$$y_{ij} - y_{aj} \neq y_{ib} - y_{ab}$$

$$\eta_{ij} - \eta_{aj} \neq \eta_{ib} - \eta_{ab},$$

which says that the net effect of changing the 1st factor from i to a is different when the 2nd factor is at j than when it is at b.

If we want to focus on direction-type information

about interactions, we are likely to do well with crossed comparisons—almost certainly better than with columns of the interaction, if for no other reason than clarity and focus of interpretation.

The simplest—and, presumably, most reasonable—way to proceed is to define

$$\text{birange} = \max_{i,j,a,b} (y_{ij} - y_{ib} - y_{aj} + y_{ab})$$

(which equals the maximum of the absolute value of the crossed comparisons) and to learn about, and use, the critical values of the Studentized birange. Just as for simple comparisons with the Studentized range, such a procedure will give, at a simultaneous error rate, a confidence interval for the underlying value of each and every bicomparison (crossed comparison).

Even together, conditional comparisons (for A and for B) and crossed comparisons (for A and B) do not *directly* address all contrasts, or even all comparisons, among the entries of A × B. But this need not worry us any more than the fact that, in a one-way table, the Studentized range does not *directly* address all the *contrasts* among the entries. Here, as well, it should suffice to restrict direct consideration, beyond conditional comparisons and crossed comparisons, to prespecified or exogenously selected comparisons or contrasts. And to include, indirectly, all logical consequences of the comparisons that are directly considered.

Multiplicity at a Higher Level, Still

We have so far been specific—at least by pointing toward alternatives—about multiplicity within A × B at three, or perhaps four, levels: (a) within a column of conditional effects; (b) among columns; and, perhaps, (c) between all within-column conditional comparisons and all nonconditional comparisons; and also, perhaps, (d) between all conditional comparisons and all crossed comparisons.

We need to think also about multiplicity at a still higher level. If we have six factors, so that there can be

$$6 \text{ main effects } (A, B, C, \ldots),$$

$$15 = \binom{6}{2} \text{ two-factor interactions}$$

$$(A \times B, A \times C, B \times C, \ldots),$$

$$20 = \binom{6}{3} \text{ three-factor interactions}$$

$$(A \times B \times C, A \times B \times D, \ldots)$$

and so on,

it is most unlikely that we will wish to spend 5% error rate on each main effect (total 30%), 5% on each two-factor interaction (total 75%), and so on. Some more sensible budgeting of error rate is quite likely to be needed.

Since we are discussing philosophy, and not procedure, it seems enough here to point this out.

DIVERSE EXTENSIONS

We now turn briefly to each of a selected set of diverse extensions.

Unequal Variances

Some attention has been paid to multiple comparisons where we have—either *a priori* or internally estimated—unequal variances for the candidates being compared. The hard questions seem to be: When do we know enough to use different estimated variances? And granted that we want to control some form of simultaneous error rate, do we want to: (a) keep individual error rates for comparisons of different precision the same, or (b) keep confidence intervals for comparisons of different precision of equal length? It may be that the preferred answer will be (b) for small or moderate unbalance, but (a) for large unbalance.

Largest Difference, No Larger Than

To set an outer bound on the largest underlying difference (the largest of a list of underlying comparisons) is essentially a fixed-price (e.g., null-price) problem and thus, like the confident-directions-only problem, it may have answers somewhat tighter than those that come from a bound provided by using the Studentized range. How much there is to gain here does not seem to be understood. At a different level, it does not seem clear how useful such a technology would be.

Smallest Difference, No Less Than

Similar remarks; greater uncertainty.

The Canyon Problem

If we have $y_1, y_2, \ldots, y_n$, but have *no* estimate of a relevant s^2, what can we sensibly do about keeping the $\{y_i\}$ together or separating them into groups? Can we find a reasonable facsimile of a canyon separating two mountains? Hoang and Tukey (1989) have made some progress on this problem.

DIFFERENCE FROM "THE BULK"

Cornfield and co-workers (Halperin, Greenhouse, Cornfield and Zalokar, 1955) introduced a multiple comparison procedure involving the differences be-

tween each individual result and a summary of all results. When these differences are called deviations, we are led to rely on the distribution of the Studentized maximum absolute deviation, conveniently "maxad" as a shorter label. When we think hard, and look at messy examples, we realize that we need to take both range-based and maxad-based approaches seriously.

Definiteness versus Specificity

The words "definiteness" and "specificity" seem quite similar, especially when applied to test statistics. But they can, and should, be given interpretations different enough to distinguish maxadlike and rangelike test statistics. Consider a collection of y's, estimating η's, which may as well have been measured with known variance. To ask whether we can specify, or be definite about, differences among the η's has two interpretations: (a) Can we pick out a particular y_i whose η_i appears to be different (in a particular direction) from the "bulk" of the η's? (Here "maxad" is responsive.) (b) Can we pick out two y's, say y_j and y_k, such that the sign of $\eta_j - \eta_k$ appears to be settled? (Here the Studentized range is responsive.) The first of these is more *specific*, because it focuses on (the "bulk" of differences involving) a *single* η_i; the second is more *definite*, because it focuses on *one* difference, $\eta_j - \eta_k$. In such a simple situation, we shall distinguish "definite" and "specific" in this way. In more general situations we would make analogous distinctions.

It is good to be definite. It is also good to be specific. In such a situation, we cannot have maximum efficiency for each while doing both.

In either case, we will usually want to list as many individual i's, or as many (j, k) pairs, as we can for which we are sure about sign—sure about direction. (Saying that there is one, without saying which, gets us almost nowhere, especially if we start from the well-founded position that all pairs of η's are different in some decimal place.) Giving a single illustration when we could give several is clearly wasteful.

At this point, the number of versions takes on great importance. If there are only 3 η's, it is probably better to know about $\eta_i - \eta_k$ than about η_i – (bulk). Definiteness can outweigh specificity in this case, especially since "bulk" is so unclear when there are only three versions in all.

If there are 300 versions, many of which are detectably different, matters are far from being the same. Even if one tenth of the y_i – (bulk) have confident directions, this is only a list with 30 entries. But if one tenth of the $\binom{300}{2} = 44{,}850$ different $y_j - y_k$ have confident directions, this is a list of 4485 comparisons, which can hardly convey any useful information to us.

There needs to be a crossover point between the use of the Studentized range and the use of a maxad procedure. Perhaps six versions of i is somewhere near crossover. Fifteen = $\binom{6}{2}$ (i, j) pairs seems a lot more than six i-values, so perhaps we should agree to use maxadlike test stiatistics for 6 or more versions and rangelike test statistics for five or fewer. The choice of six is not very firm, but four seems too few, and ten seems too many.

Circumstances may indeed alter cases here, but we have not thought things through adequately; we do not even understand what kinds of circumstances would matter.

Which "Bulk"?

Once we are pledged to construct a list of those y_i – (bulk) that are of confident sign, we have to deal with a problem of confusion. We dare not take "bulk" to be the mean of the y's, since situations like

$$-10, -9, -8, -7, -6, -5, -4, -3, -2, -1,$$
$$0, 1, 2, 1000, 14{,}050$$

with a mean $\bar{y}$ of 1000, would give, for any reasonable standard error (say, between 2 and 200), 13 negative and apparently significant $y_i - \bar{y}$, 1 zero, and 1 very positive. A conclusion that all but the 1000 are different from the "bulk" would be at best useless and almost certainly misleading.

We may dare to take the median of the y's as the "bulk," though situations where 40% of the y's are obviously high, but none are obviously low, leave us somewhat worried about using the 50% point of the original collection to describe a bulk ranging from 0% to only 60%. Using a (one-step) biweight instead of the median may, however, be an adequate bandage for such a small wound.

A more careful procedure would begin by defining subgroups by applying only the gap portion of one of Welsch's (1977) procedures involving gaps and stretches and would then continued with a maxad approach using *subgroup* medians *within* each subgroup for the "bulk" and P-values based on Bonferroni and the number of subgroups. Though the overall P-value of this procedure deserves study, it does not seem likely to be far enough above nominal for us to worry.

Notice that, in the absence of large gaps, when using the median is most reasonable, this procedure will reduce to maxad using "y_i – median." Further, if there are detectable gaps, its results will combined specificity and definiteness by telling

us, clearly and firmly, about confident directions between whole subgroups.

An Historical Note

Besides the early, unfollowed-up remarks of Hotelling (1931), there were at least three independent and nearly simultaneous incursions into the thicket of multiple comparisons procedures. Scheffé's explication of the F test (Scheffé 1953) was compatible with inadequately founded ideas of equal justice, but it proved far too wasteful for simple comparisons, the contrasts of greatest interest and importance.

The Studentized-range approach, arising from a request (from the Baltimore Section of ASQC, probably in the fall of 1951) to the author for a talk on "Industrial uses of the range," focused its attention on simple comparisons, had a simple theory, and used available tables.

The work of Cornfield and co-workers (Halperin, Greenhouse, Cornfield and Zalokar, 1955) initiated the maxad approach, which now seems to be—or to be part of—the preferred approach, once we face more than a few (or perhaps more than several) versions. It did not receive as much attention at the time, for which some or all of the following may have been important reasons:

(a) Tabulation of the critical values of the test statistic was not easy; the original paper offered bounds rather than values.

(b) Coming from statisticians at the National Institutes of Health, it may have been regarded as suited to rather special problems, like identifying a few apparently effective drugs in a drug-screening program. (The selection literature did—and still does—favor two- or more-stage experiments, where the initial stage will not reach significance!)

(c) Techniques and ideas such as those mentioned above—use of medians or biweights instead of means, combination with Welsch's subgrouping techniques—were either not familiar or not available.

(d) Except in special fields, experiments with many versions were not common, and extensions to two-way tables had not been dreamed of.

(e) We did not stop to think hard enough (in part, perhaps, because of the distraction of Duncan's proposals (e.g., 1955, 1965), which amounted to "talking 5%" while using more than 5% simultaneous).

Times have changed, and we should be prepared to change our ideas with them.

ACKNOWLEDGMENTS

The author acknowledges with pleasure, as each reader should also, the improvements in clarity due to the comments of David C. Hoaglin and an anonymous reviewer. Unclarities and errors remaining are, of course, the writer's sole responsibility. This article is based on the 1989 Miller Memorial Lecture at Stanford University, prepared in connection with research at Princeton University sponsored by the Army Research Office (Durham), DAAL03-86-K-0073 and DAAL03-88-K-0045.

REFERENCES

BEGUN, J. M. and GABRIEL, K. R. (1981). Closure of the Newman-Keuls multiple comparisons procedure. *J. Amer. Statist. Assoc.* **76** 241-245.

BRAUN, H. I. and TUKEY, J. W. (1983). Multiple comparisons through orderly partitions: The maximum subrange procedure. In *Principals of Modern Psychological Measurement: A Festschrift for Frederic M. Lord* (H. Wainer and S. Messick, eds.) 55-65. Erlbaum, Hillsdale, N.J.

DUNCAN, D. B. (1955). Multiple range and multiple F tests. *Biometrics* **11** 1-42.

DUNCAN, D. B. (1965). A Bayesian approach to multiple comparisons. *Technometrics* **7** 171-222.

FISHER, R. A. (1926). The arrangement of field experiments. *J. Min. Agric. G. Br.* **33** 503-513. (Also pages 83-94 in Volume 2 of his collected papers.)

HALPERIN, M., GREENHOUSE, S. W., CORNFIELD, J. and ZALOKAR, J. (1955). Tables of percentage points for the Studentized maximum absolute deviate in normal samples. *J. Amer. Statist. Assoc.* **50** 185-195.

HOANG, T. and TUKEY, J. W. (1989). Procedures for separations within batches of values, I. The orderly tool kit and some heuristics. Technical Report 293, Dept. Statist., Princeton Univ.

HOCHBERG, Y. and TAMHANE, A. C. (1987). *Multiple Comparison Procedures*. Wiley, New York.

HOTELLING, H. (1931). The generalization of Student's ratio. *Ann. Math. Statist.* **2** 360-378.

KAFADAR, K. K. (1982). Using biweight M-estimates in the two-sample problem. Part 1: Symmetric populations. *Comm. Statist. Theory Methods* **11** 1883-1901.

KUHN, T. S. (1970). *The Structure of Scientific Revolutions*, 2nd ed. Foundations of the Unity of Science Series **2** (2) Univ. Chicago Press.

MILLER, R. G. (1966). *Simultaneous Statistical Inference*, McGraw Hill, New York. (2nd ed., Springer, New York, 1981.)

PERITZ, E. (1970). A note on multiple comparisons. Unpublished manuscript, Jerusalem, Hebrew University. (Cited by Begun and Gabriel, 1981).

RAMSEY, P. H. (1981). Power of univariate pairwise multiple comparison procedures. *Psych. Bull.* **90** 352-366.

SCHEFFÉ, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika* **40** 87-104.

WELSCH, R. E. (1977). Stepwise multiple comparison procedures. *J. Amer. Statist. Assoc.* **72** 566-575.