OPTIMALITY RESULTS IN MULTIPLE HYPOTHESIS TESTING

JULIET POPPER SHAFFER

ABSTRACT. Multiple hypothesis testing occurs in a vast variety of fields and for a vast variety of purposes. Optimality results are relatively sparse in this area compared to results for tests of individual hypotheses. This paper restricts consideration to cases in which a finite number of parameters are involved, in which conclusions are desired for each parameter separately, and in which directional inference may or may not be involved. The paper does not deal with optimal design, with tests of composite hypotheses without further resolution, nor with sequential analysis or ranking and selection procedures. It is primarily a historical survey, concentrating on work of Erich Lehmann, but relating his optimality results to more recent developments, primarily in stepwise testing.

1. INTRODUCTION

Testing of more than one hypothesis simultaneously is widely practiced in many fields and for many purposes. While theory and methods in this area originally arose in connection with relatively small numbers of treatments being evaluated or compared, more recently there has been a great increase in applications to situations in which massive numbers of hypotheses are being considered jointly, such as in large surveys (for example, the National Assessment of Educational Progress–see Beaton and Zwick, 1992), signal compression, microarray analysis, and astronomy. In some of these situations, the number of hypotheses ranges into the tens of thousands and higher.

An alternative way of looking at multiple hypothesis testing is as a multiple decision problem. It turns out to be useful to consider such problems from both multiple testing and multiple decision perspectives, and sometimes in fact to combine these two perspectives. These dual perspectives and combinations will be illustrated.

There is some overlap between this paper and Chapter 11 of Hochberg and Tamhane (1987) on optimal procedures, but the present paper also covers some material, related to stepwise hypothesis testing procedures, not covered in that source. Points of intersection will be noted in the ensuing discussion.

When considering multiple testing problems, the concern is with Type I errors when hypotheses are true, and Type II errors when they are false. The evaluation of procedures is based on criteria involving balance between these errors, or special attention to Type I errors. When considering these problems from a multiple decision standpoint, losses are attached to each incorrect decision, and some method of minimizing losses is derived. Bayesian methods incorporate prior distributions on parameters in addition to other assumptions in each of these approaches.

In the hypothesis testing approach, it is important to specify the type of error control of interest. The criterion that has been used most widely is control of the familywise error rate ($FWER$), following the terminology formulated by Tukey (1952, 1953). This is defined as the probability of one or more errors in a family of hypotheses under consideration; alternatively, as the probability of one or more incorrect decisions. With small experiments, this rate was sometimes referred to in the past as the experimentwise error rate, since it was applied over all hypotheses being tested (see Ryan, 1959). With the increasing application of these ideas to larger and more complex investigations with many hypotheses, it is appropriate to use the more general term familywise error rate, since error control of this type may be too conservative when applied to all hypotheses to be tested. There is no clear criterion on how to define families over which errors should be controlled; for example, in a multifactor experiment, should errors be controlled at a specified level within each factor? How about interactions? See Westfall and Young (1993, Chapter 7) for an interesting discussion of this issue with examples.

An alternative criterion, the per-family error rate ($PFER$) (Tukey, 1952) is control of the expected number of falsely rejected hypotheses. When the probability of rejecting any single hypothesis is small, and the test statistics are independent, the $PFER$ is only slightly greater than the $FWER$ for most common testing procedures, and it is conventional to use the same upper limit under both error criteria. With dependent test statistics the difference can be considerable, making the $PFER$ control at a given level substantially more conservative than the $FWER$ control at that level. Note that the $PFER$, in distinction to the $FWER$, can be expressed in an additive form over the family of hypotheses, using an indicator variable for false rejection of each hypothesis. This makes the $PFER$ easier to deal with statistically, as will be illustrated.

Finally, a more recently-introduced criterion, control of the false discovery rate (FDR), will be discussed briefly.

2. The One-sided Problem

A simple example of the dual way of considering multiple problems relates to one-sided tests of two hypotheses: $H_i : \theta_i = 0$ with the alternatives $A_i : \theta_i > 0, i = 1, 2$. Looked at as hypothesis tests, the test statistics and critical values must be specified for the two tests. Considered as multiple decisions, probabilities of the four decision regions $(\theta_1 = 0, \theta_2 = 0), (\theta_1 = 0, \theta_2 > 0), (\theta_1 > 0, \theta_2 = 0)$, and $(\theta_1 > 0, \theta_2 > 0)$ must be specified. Alternative formulations replace "=" with "$\leq$" in H or replace ">" with "$\geq$" in A, with the obvious changes in the decision regions. With continuous variables, these different formulations do not affect the procedure. Of course the alternative can be in the negative rather than the positive direction, again with the obvious changes. If $\theta_i = 0$ is considered impossible for all i, the hypotheses can be expressed in the form $H_i : \theta_i < (>)0$ with the alternative $A_i : \theta_i > (<)0$. This formulation does result in a change in procedure (see Jones and Tukey, 2000, Shaffer, 2002). Further discussion on this point is in the later section on the two-sided problem.

When k parameters are involved, one-sided tests can be similarly formulated. Without loss of generality, it will be assumed that $\theta_0 = 0$ in all hypotheses.

$H_i : \theta_i = 0(or \leq 0), i = 1, \ldots\ldots, k$ with alternatives $A_i : \theta_i > 0$.
As a decision problem, this involves 2^k possible decisions.

Two early articles with optimal results for one-sided tests in a multiple setting appeared in the same issue of the 1952 Annals of Statistics. Both formulate the problem as multiple decisions, but use Neyman-Pearson hypothesis testing concepts as well. Both deal with one-sided hypotheses as formulated above.

The paper by Paulson (1952) is more limited in application than the Lehmann paper, and is discussed more technically in Hochberg and Tamhane (1987). It assumes samples of equal size n from k categories, all observations normally distributed, and the means of all categories equal except possibly for one higher mean. This is an example of a slippage problem, an area that has been studied with many variations since a paper by Mosteller (1948), and which has applications to outlier detection. For a general treatment of the area, see Schwager (1985). Paulson's paper was the first to treat a slippage problem from an optimal point of view. Although the present paper does not deal further with such problems, the Paulson result is included for historical purposes.

Paulson limited procedures to those with two characteristics:

(1) symmetry: The probability of correctly concluding that category i is best is the same for each $i, i = 1, ..., k$,

and

(2) location-scale invariance: Subtracting the same constant from each mean and/or multiplying each mean by the same constant does not change the procedure.

The null hypothesis is that all means are equal, (without loss of generality equal to zero) and the alternative is that all are equal except one, which is larger by an unknown amount Δ, the same value for each category. This restricts the decision space to $k+1$ decisions, $D_i, i = 0, 1, \ldots, k$, where D_0 is accepting the null hypothesis, and $D_i, i = 1, \ldots, k$, is deciding category i is best.

Paulson uses the Neyman-Pearson approach in setting the probability of accepting the null if it is true at $1 - \alpha$. Although no prior probability is then assigned to the complete null hypothesis, the Bayesian approach is used in assigning the same prior probability of being best to each category. Given these restrictions, the probability of making the correct decision (i.e. choosing the category with the larger mean) is maximized. The optimum procedure is: if

$$\frac{n(\overline{x}_M - \overline{x})}{\sqrt{\sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - \overline{x})^2}} > \lambda_\alpha$$

select D_M; otherwise select D_0,

where $\overline{x}_M$ is the maximum of $\overline{x}_1, \ldots, \overline{x}_k$, $\overline{x}$ is the mean of all groups combined, and λ_α is a constant depending only on α.

Note that by restricting the situation to only one category being larger than all the others, which are all equal, the number of decisions to be considered and the complexity of the distributions is greatly reduced, compared to more general alternatives. While it seems possible to extend the class of distributions from normal to other independent distributions, the restriction to a single deviation from a common mean would be harder to overcome. [Truax (1953) used the same approach and formulation to find an optimal procedure for a similar slippage problem on normal variances.]

In the same issue of the 1952 Annals, there is a paper by Lehmann entitled "Testing multiparameter hypotheses." This paper deals with multiple parameters but most of the paper does not treat multiple comparisons. Instead, it treats composite hypotheses and alternatives of the forms

(a) $H_i : \theta_i \leq \theta_i^*, i = 1, \ldots, k$ with the alternative that at least one $\theta_i > \theta_i^*$, and

(b) $H_i : \theta_1 \geq \theta_1^*$ or $\ldots$ or $\theta_k \geq \theta_k^*$ with the alternative that all $\theta_i < \theta_i^*$.

These two situations essentially reverse hypothesis and alternative, but in each case there is one hypothesis and one alternative. However, one small section of about one page deals with a multiple testing situation. The discussion is limited to $k = 2$, with random variables called X_1 and X_2. The paper considers a class of distributions satisfying certain regularity conditions, generalizing well beyond independent and identically distributed normal variables, and including the situation considered by Paulson when $k = 2$. The marginal distribution of X_i is assumed to depend only on θ_i. Without loss of generality, let $\theta_{i0} = 0$ for $i = 1, 2$.

The paper restricts attention to one-sided hypotheses, as noted above. The null hypotheses then are $H_1 : \theta_1 \leq 0$, $H_2 : \theta_2 \leq 0$. Two restrictions for the class of tests are:

(a) that they be symmetric in X_1 and X_2 (this assumption may not be necessary in this context), and

(b) that they be monotone in the regions where both hypotheses are rejected and where both are accepted, i.e. that if x_1 and x_2 are in the joint rejection region, and $x_1' \geq x_1, x_2' \geq x_2$, then x_1' and x_2' are in the rejection region. Similar natural monotone restrictions are imposed in the opposite direction in the joint acceptance region, and in single-rejection regions.

In decision-theoretic terms, the four parameter regions of interest are
$\omega_0 : \theta_1$ and $\theta_2 \leq 0$ (both hypotheses are true),
$\omega_1 : \theta_1 > 0 \geq \theta_2$ (H_1 is false, H_2 is true),
$\omega_2 : \theta_2 > 0 \geq \theta_1$ (H_2 is false, H_1 is true) and
$\omega_{12} : \theta_1, \theta_2$ both > 0 (both hypotheses are false).
The corresponding correct decisions are d_0, d_1, d_2, and d_{12}. The complements of each set ω will be called $\overline{\omega}$, and the set of all decisions except d will be called $\overline{d}$.

An optimal procedure is derived as follows, subject to satisfaction of (a), (b), and some regularity conditions.

1. Set the familywise error rate (i.e. the probability of rejecting one or more true hypotheses) less than or equal to a preassigned value α. In decision-theoretic terms,
$$P(\overline{d}_0|\omega_0),\ P[(d_1 \bigcup d_{12})|\theta_1 \leq 0],\ P[(d_2 \bigcup d_{12})|\theta_2 \leq 0],\ \text{and } P(d_{12}|\overline{\omega}_{12}) \leq \alpha.$$
2. Maximize $min\ P(\overline{d}_0|\overline{\omega}_0)$.
3. Maximize $min\ P(d_{12}|\omega_{12})$.

This results in the following test procedure, described for simplicity under the assumption that X_1 and X_2 are identically distributed.
Decide d_0 if $max(x_1, x_2) \leq a$ where $P_{0,0}[max(x_1, x_2) \leq a] = \alpha$.
Decide d_{12} if $x_1 > a,\ x_2 > b$ or $x_1 > b,\ x_2 > a$,
where $P(X_1 > b|\theta_1 = 0) = P(X_2 > b|\theta_2 = 0) = \alpha$.
Decide d_i if $X_i > a, X_{i'} < b$.

Lehmann formulated this procedure in multiple-decision terms. A reformulation as a multiple testing problem shows that it is a generalization for two hypotheses of a familiar step-down test (Holm, 1979).

Step-down tests can be contrasted with single-step tests. With single-step tests, each hypothesis is tested using a criterion for the test statistic that may depend on the number of hypotheses k but is independent of the realized values of all test statistics for hypotheses $H_j, j \neq k$. Step-down tests are a subclass of the class of stepwise tests; the latter are defined as procedures in which there is a single test statistic for each hypothesis but its critical value depends on the realized values of test statistics for other hypotheses in a specified pattern.

Some of the earliest stepwise test procedures, although they are not usually designated by that name, are the multiple range procedures developed for comparing parameters (Tukey, 1953, Duncan, 1961, Ryan, 1959, 1960). Lehmann and Shaffer (1979) defined an optimality criterion and gave optimal choices of critical values for these methods. However, the methods themselves are not optimal for the problems being considered, so they will not be treated further here.

The stepwise tests considered in this paper are based on p-values; for other types of stepwise procedures, valid using resampling methods, see Westfall and Young (1993). The step-down procedures have the following form: Let p_i be the p-value for the test of H_i, and reorder the p-values and associated hypotheses so that $p_1 \leq \ldots \leq p_k$. Suppose i is the smallest integer from 1 to k for which $p_i \geq \alpha_i$. If there is no such integer, reject all hypotheses. Otherwise reject $H_1, \ldots, H_{i-1}$, and accept $H_i, \ldots, H_k$. The α_i's are selected to satisfy the specified error control conditions.

Thus, in modern terms, the procedure in Lehmann (1952) described above is a step-down testing procedure. Table 1 specifies the test and the probabilities of the decision regions at $\theta_i = 0, i = 1, 2$, for $k = 2$, independent test statistics, compared with those probabilities for a single-stage test with the same $FWER$.

Step-up testing procedures are in a sense the reverse of step-down testing procedures. Using the notation above, let p_i be the largest integer from 1 to k for which $p_i \leq \alpha_i$. If there is no such integer, accept all hypotheses. Otherwise reject $H_1, \ldots, H_i$ and accept $H_{i+1}, \ldots, H_k$. The α_i's are selected to satisfy the specified error control conditions.

Suppose the requirements that result in the step-down procedure of Lehmann (1952) are modified by reversing steps 2 and 3. The requirements then, each subject to satisfaction of the previous ones, are:

1. Set the familywise error rate (i.e. the probability of rejecting one or more true hypotheses) less than or equal to a preassigned value α. In decision-theoretic terms, $P(\bar{d}_0|\omega_0)$, $P[(d_1 \bigcup d_{12})|\theta_1 \leq 0]$, $P[(d_2 \bigcup d_{12})|\theta_2 \leq 0]$, and $P(d_{12}|\bar{\omega}_{12}) \leq \alpha$.

2. (formerly 3) Maximize $min\ P(d_{12}|\omega_{12})$.

3. (formerly 2) Maximize $min\ P(\bar{d}_0|\bar{\omega}_0)$.

This results in the following optimal step-up multiple test procedure for two hypotheses:

Decide d_{12} if $min(x_1, x_2) \geq b'$, where
$P(X_1 > b'|\theta_1 = 0) = P(X_2 > b'|\theta_2 = 0) = \alpha$.

TABLE 1. The one-sided optimal step-down procedure for one-sided tests, Lehmann (1952), $FWER \leq \alpha, k = 2$, independent test statistics, comparison with single-step $PFER$-controlling test. Decision region probabilities are for $\theta_1, \theta_2 = 0$. Let $p_i = Prob(T_i > t_i)$, ordered so that $p_1 \leq p_2$, where t_i is the observed value of the statistic T_i for testing H_i.

Let $\beta = 1 - (1 - \alpha)^{(0.5)}$

Single-step test

Reject H_i if $p_i \leq \beta, i = 1, 2$

Step-down test

Step 1: Reject H_1 if $p_1 \leq \beta$, otherwise accept both H_i

Step 2: Reject H_2 if H_1 rejected and $p_2 \leq \alpha$

Single-step test	Step-down test
$P(d_0) = (1 - \beta)^2 = 1 - \alpha$	$P(d_0) = (1 - \beta)^2 = 1 - \alpha$
$P(d_1) = P(d_2) = \beta(1 - \beta)$	$P(d_1) = P(d_2) = \beta(1 - \beta)^2$
$P(d_{12}) = \beta^2$	$P(d_{12}) = 3\beta^2 - 2\beta^3$

Decide d_1 if $x_1 > a', x_2 < b'$; d_2 if $x_2 > a', x_1 < b'$, where
$P(X_1, X_2 > b'|\theta_1 = 0) + P(X_1 > a', X_2 < b'|\theta_1 = 0) =$
$P(X_1, X_2 > b'|\theta_2 = 0) + P(X_2 > a', X_1 < b'|\theta_2 = 0) = \alpha.$

It seems likely that this approach can be generalized to obtain optimality of step-up and step-down methods, and possibly more general stepwise up-down methods (Tamhane, Liu, and Dunnett, 1998) for more than two hypotheses, by generalizing the symmetry and monotonicity assumptions and adjusting the priority ordering of the requirements, as follows:

Assume k hypotheses $H_i : \theta_i \leq 0$, as before, with statistics $X_i, i = 1, \ldots, k$. There are now 2^k regions in the parameter space and 2^k associated decisions.

1. Set the familywise error rate (i.e. the probability of rejecting one or more true hypotheses) less than or equal to a preassigned value α. This requirement is easily generalized.

For step-down procedures:

2. Maximize $min\ P(\overline{d}_0|\overline{\omega}_0)$. This is also easily generalized and results in: Decide d_0 if $max(x_i) < a$.

For step-up procedures:

2. Maximize $min\ P(d_{12\ldots k}|\omega_{12\ldots k})$. This also generalizes and results in: Decide $d_{12\cdots k}$ if $min(x_i) > b'$.

More requirements are necessary to specify the divisions of the sample space among the remaining decisions, and to specity the divisions for more general up-down methods. The monotonicity restriction seems to be very reasonable when the distributions have monotone likelihood ratios, but recently there has been some development of tests that are more powerful when that restriction is removed.

In Lehmann (1952), an example of an "unreasonable", more powerful, nonmonotone rejection region for testing the hypothesis

$H_0 : \theta_1 \leq 0$ or $\theta_2 \leq 0$ with the alternative $\theta_1 > 0$ and $\theta_2 > 0$

was described. Let d_0 here mean accepting H_0 and d_{12} mean rejecting H_0.

Under the monotone restriction, and the Type I error restriction
$P(d_0|H_0) \geq 1 - \alpha$, the procedure that maximizes the minimum power is
Decide d_{12} if $min(x_i) > b'$, where b' is chosen so that $P(X_1 > b'|\theta_1 = 0) = P(X_2 > b'|\theta_2 = 0) = \alpha$.
Note that the minimum power is $< \alpha$: e.g. if X_1 and X_2 are independent, the minimum power is α^2.

Lehmann described a general way of formulating nonmonotone procedures with greater power than the monotone procedure for testing this hypothesis, in which the monotone rejection region is increased by a union of one-sided polygonal regions. A number of papers have considered this and other methods for obtaining more-powerful nonmonotone procedures for these and for somewhat more general hypotheses. Berger (1989) has explored this area most extensively, adopting the extension of the monotone region formulated by Lehmann for testing one-sided hypotheses and generalizing the approach to tests of two-sided hypotheses. Zelterman (1990) derived a locally-most-powerful nonmonotone procedure which is somewhat like a smoothed version of the Berger procedure, with a curved regection region.

Perlman and Wu (1999) attacked these nonmonotone tests on intuitive grounds. They support the likelihood ratio principle, which yields monotone tests, but, as they point out, even the LR principle does not always result in intuitively desirable test procedures. It seems unlikely that *any* optimality principle will always yield sensible-appearing procedures. At any rate, the monotonicity requirement seems as reasonable as any other criterion that might be applied.

The optimal procedures considered above control the $FWER$. The $PFER$, as noted above, is technically easier to work with than the $FWER$. Results in Lehmann (1957a,b) and Spjøtvoll (1972) based on $PFER$ control, will be discussed in the context of two-sided tests, but they apply also, as in the two-sided case, to yield single-step one-sided tests.

3. The two-sided problem

3.1. **Directional and nondirectional errors.** The treatment thus far has been for the one-sided problem–testing the hypotheses $\theta_i \leq 0$ vs. the alternative $\theta_i > 0$. Another major interest in testing hypotheses concerning multiple parameters is the two-sided problem. The hypotheses in that case are usually formulated as

$H_i : \theta_i = \theta_{i0},\ i = 1, \ldots, k.$

Without loss of generality, it will be assumed that $\theta_{i0} = 0$ for all i.

The interpretation of acceptance of a two-sided hypothesis H_i has been a subject of some dispute. From one point of view, it can be interpreted as a decision to behave as though $\theta_i = 0$. From an inference point of view, it seems more reasonable to regard acceptance as uncertainty about the value of θ_i. Some but not all investigators consider any exact value of θ_i impossible. Others interpret the hypothesis as meaning that θ_i lies in some infinitesimal region around zero, to be treated as zero (Scheffe, 1970). In any case, if the two-sided hypothesis H_i is rejected, it is often important to know whether θ_i is positive or negative.

When direction is important, there are a number of different possible approaches to the treatment of errors. The first is a reformulation of the standard approach to testing H_i as a joint test of two hypotheses:

D1. If $\theta_i = 0$, any decisions involving $\theta_i > 0$ or $\theta_i < 0$ are errors. Thus, there is a discontinuous increase in the probability of error at $\theta_i = 0$. The appropriate formulation is

$H_{i1} : \theta_i \leq 0, A_{i1} : \theta_i > 0,\ H_{i2} : \theta_i \geq 0, A_{i2} : \theta_i < 0.$

D2. If $\theta_i = 0$ is considered impossible, and Scheffe's alternative of an infinitesimal region around zero is rejected, then only directional errors are of interest. The appropriate formulation is

$H_{i1} : \theta_i < 0, A_{i1} : \theta_i > 0,\ H_{i2} : \theta_i > 0, A_{i2} : \theta_i < 0.$

D3. If $\theta_i = 0$, the decision is unimportant. Only directional errors are counted. Thus, there is a discontinuous drop (to zero) in the probability of error at $\theta_i = 0$. The formulation would be the same as in 1, but with no penalty for any decision if $\theta_i = 0$.

D4. If $\theta_i = 0$, this is included with either $\theta_i < 0$ or $\theta_i > 0$, depending on the consequences for practice. Thus, the probability of error is continuous at $\theta_i = 0$. The appropriate formulation is

$H_{i1} : \theta_i > (\geq) 0, A_{i1} : \theta_i \geq (>) 0,\ H_{i2} : \theta_i > (\geq) 0, A_{i2} : \theta_i \leq (<) 0.$

The two-sided problem can be thought of alternatively as a three-decision problem: Deciding θ_i is positive, negative, or possibly zero. If positions D2 or D4 are taken with respect to $\theta_i = 0$, it would seem that a two-decision problem results. However, Bohrer (1979) showed that if the error probability is to be controlled at any level below 0.5, there must be a possible third decision: The sign of θ_i is indeterminate. Thus, in all cases in which direction is of interest, three decisions must be permitted, as in the Lehmann (1957b) formulation.

Shaffer (2002) uses the terms nondirectional error to refer to rejections of true null hypotheses (even if directional conclusions are attached), and directional error to refer to false directional conclusions (stating $\theta_i > 0 (< 0)$ if in fact $\theta_i < 0 (> 0)$. In view of interpretation D4 above, the definition of directional error should be modified to refer to rejecting one correct hypothesis of the pair of directional hypotheses when the other is incorrect, and nondirectional error to refer to rejecting the single hypothesis $H_i : \theta_i = 0$ for each i, or the pair of hypotheses in D1, when $\theta_i = 0$. One might then be interested in controlling nondirectional error only, directional error only, or both types of errors, called combined errors by Shaffer (2002). Some common stepwise multiple testing methods for the hypotheses D1 do not necessarily control combined errors (Shaffer, 1980, Liu, 1997) while maintaining the nominal familywise error rate ($FWER$). In other cases, it is not known whether directional conclusions are permissible without violating the nominal $FWER$. For a review of the research on this issue, see Finner (1999).

Lehmann (1957a) considered the set of directional hypotheses (H_{i1}, H_{i2}, $i = 1, \ldots, k$). For $k = 2$, for example,
in testing the four hypotheses

$$H_{11} : \theta_1 \leq 0,\ H_{12} : \theta_1 \geq 0;\ H_{21} : \theta_2 \leq 0,\ H_{22} : \theta_2 \geq 0,$$

the parameter space can be partitioned into nine regions corresponding to all combinations of the possible situations for each θ_i: that it equals 0, is < 0, or is > 0. For $k > 2$, there will be 3^k such regions in the parameter space. Decisions are specified by dividing the sample space into 3^k regions corresponding to those in the parameter space.

However, under this formulation, some of these regions will be empty in some important types of problems. For example, for $k = 3$, suppose there are samples from three populations and the hypotheses are:

$$H_{11} : \theta_1 - \theta_2 \leq 0, H_{12} : \theta_1 - \theta_2 \geq 0,$$
$$H_{21} : \theta_1 - \theta_3 \leq 0, H_{22} : \theta_1 - \theta_3 \geq 0,$$
$$H_{31} : \theta_2 - \theta_3 \leq 0, H_{32} : \theta_2 - \theta_3 \geq 0.$$

There is no set in the parameter space corresponding to the decisions

$$\theta_1 - \theta_2 = 0,\ \theta_2 - \theta_3 = 0,\ \theta_1 - \theta_3 \neq 0.$$

Nevertheless, those decisions are possible when using the tests that are optimal in a general context.

In a later paper Lehmann (1957b) adopted an alternate formulation that solved this problem. Instead of partitioning the parameter space, the decision to accept a hypothesis was interpreted as providing no information on the value of the parameter involved. So instead of the impossible conclusion

$$\theta_1 = \theta_2,\ \theta_2 = \theta_3,\ \theta_1 < \theta_3,$$

the conclusion would be

$-\infty < \theta_1 - \theta_2 < \infty$, $-\infty < \theta_2 - \theta_3 < \infty$, $\theta_1 < \theta_3$.

Then all decisions are mutually consistent.

3.2. **Per-family error control.** Lehmann (1957b) approached the problem from a purely decision-theoretic point of view, and found an optimal procedure given control of $PFER$.

The pair of directional hypotheses (1) is tested for each parameter. A loss function is defined as follows: Loss $=0$ for a correct decision, a for rejecting a hypothesis if it should be accepted, and b for accepting a hypothesis if it should be rejected. Thus, both directional and nondirectional errors are penalized. Adding the losses over the two decisions in each pair gives the loss table for comparing that pair; adding the losses over the decisions involving all pairwise comparisons gives the loss table for the set of comparisons.

If a (the loss for false rejection) and b (the loss for false acceptance) are the same for each hypothesis, and if each hypothesis is tested with a uniformly most powerful unbiased test at level $b/(a+b)$, the natural resulting multiple procedure is unbiased (in the sense of Lehmann, 1951) and has uniformly minimum risk among unbiased procedures.

Spjøtvoll (1972), also in the context of a more general formulation, noted that optimality results for individual hypotheses would be desirable as well as the optimality results based on the global criterion adopted by Lehmann. He adopted $PFER$ control, and gave conditions and stated values under which single-stage two-sided testing procedures maximized the minimum individual power and the minimum average power of the tests. His results apply both to one-sided and two-sided tests, as do Lehmann's (1957a,b). Although he formulated the pair of two-sided hypotheses as

$H_i : \theta_i = 0$ vs. $A_i : \theta_i > (<)0$

rather than the more general

$H_i : \theta_i <= (>=)0$ vs $A_i : \theta_i > (<)0$,

his results, in slightly less generality, would apply to the latter formulation as well. For a simple example, if the k pairwise hypotheses specify that the expected values of k independent normal variables with equal variances equal zero, t tests maximize the minimum individual power and the minimum average power against a common alternative Δ.

Bohrer (1979), under the formulation D4 above, uses Spjøtvoll's (1972) result and takes limits as the hypothesized values approach zero to prove, for normally-distributed variables, that two-sided t-tests of k hypotheses, each at the same level, maximize the minimum probability of correct classification of the θ_i's as positive or negative (zero included in one of these), under the assumption that the expected number of misclassifications (the $PFER$) is $\leq \alpha$, for some α. The results apply also under formulations D2 and D3. Thus, Bohrer considers directional errors only.

The results of Lehmann (1957b), Bohrer (1979), Spjøtvoll (1972), and some Bayesian results to be discussed below, all involve additive error criteria, and all show optimality of single-stage tests given such criteria. All of them are discussed more technically and in greater generality in Hochberg and Tamhane (1987).

3.3. **Familywise error control.** The results for $FWER$ control in the two-sided case, as in the one-sided case, demonstrate that single-stage tests are not optimal under that error criterion.

In Lehmann (1952) the criterion was control of the $FWER$. This is probably the most frequent criterion in current practice. With $PFER$-control, optimal methods are generally single-stage methods that are unable to take advantage of the potential increases in power based on rejection of other hypotheses in the family under consideration. Stepwise methods that control $FWER$ but not $PFER$ are more likely to reject false hypotheses than single-step methods with the same $FWER$ control, with the advantage of increasing power when some hypotheses are false. The stepwise test procedure of Holm (1979), for example, which is completely general, improves on the corresponding single-step Bonferroni procedure. Results of Hochberg (1987) show superiority of stepwise over single-step procedures in general.

Lehmann (1952), discussed above, dealt with a set of one-sided hypotheses and control of the familywise error rate. Optimality in the two-sided case is more complex under this nonadditive criterion.

Such optimality has been investigated by Bohrer (1982), Bohrer and Schervish (1980), Hochberg (1987), and Hochberg and Posner (1986). They adopted formulation 4 above, treating $\theta_i = 0$ as equivalent in decision consequences to either $\theta_i < 0$ or $\theta_i > 0$. Their results apply equally to formulations D2 and D3, but not to formulation D1.

These papers all require procedures to satisfy a generalization of the symmetry requirement stated for the one-sided tests in Lehmann, and to what is called upper convexity, which is a generalization of Lehmann's one-sided monotonicity assumption–essentially monotonicity in each of the 2^k parameter regions in which all hypotheses $\theta_i = 0$ are rejected. Hochberg and Posner generalize this further to include monotonicity in the parameter regions in which only some hypotheses are rejected. This more general criterion can be stated as follows, and implies the Bohrer-Schervish condition:

Let $D_i = 1, -1$, or 0 as θ_i is designated positive, negative, or indeterminate, respectively. Then if $D_i = 1(-1)$ given $x_i, i = 1, \ldots, k$, for any i, $D_i = 1(-1)$ given $c_i x_i, i = 1, \ldots, k$ when $c_i \geq 1$ for all $i = 1, \ldots, k$.

Interestingly, only Hochberg (1987) cites the Lehmann (1952) paper or notes the similarity of the criteria to those in that paper. Bohrer (1982), Bohrer and Schervish (1980), Hochberg (1987), and Hochberg and Posner (1986) consider optimality of various kinds when the parameter vector approaches zero, which they call local optimality.

Results are given for normally-distributed test statistics, although presumably they could be generalized to apply to other distributions. The simplest results apply when the test statistics are independent, in which case, under very general assumptions, knowledge of the distribution is needed only for determining critical values for the relevant statistics. The optimality criterion that receives the greatest attention is maximization of the expected number of correct decisions as the parameter vector approaches zero. Sections 3.3.1 and 3.3.2 give results for this optimality criterion.

3.3.1. *Independent test statistics, $k = 2$.* For the main part, results are limited to the case $k = 2$. When $FWER \leq 1/3$, Bohrer and Schervish (1980) obtain a two-sided step-up procedure as one of a class of optimal procedures, while Hochberg and Posner (1986) show that the step-up procedure Bohrer and Schervish selected as intuitively most reasonable in that class is the unique optimal procedure in the limit as $(\theta_1, \theta_2) \to 0$ where $\theta_1 \geq 0, \theta_2 \geq 0$ and at least one $\theta_i > 0$. (In the course of that proof they give a rule for maximizing the expected number of correct decisions for all $\theta_1 \geq 0, \theta_2 \geq 0$.) By symmetry, the results apply with obvious directional changes to $\theta_1 \geq 0, \theta_2 \leq 0$, etc. The optimal procedure is a two-sided version of the general step-up procedure defined earlier, allowing directional inferences in both positive and negative directions. The specific designation of the procedure for independent test statistics is given in Table 2.

TABLE 2. The optimal two-sided step-up procedure for two-tailed tests, Bohrer and Schervish (1980), Hochberg and Posner (1986), $FWER \leq \alpha$ for directional errors, $k = 2$, independent test statistics, comparison with single-step test controllng $PFER$ for directional errors. Assume $\theta_1, \theta_2 > 0$. Decision region probabilities are limits as $\theta_1, \theta_2 \rightarrow 0$. For simplicity, assume test statistics T_i for testing $H_i, i = 1, 2$ with limiting distributions symmetric around 0. Let $p_i =$ limiting Prob $(|T_i| > |t_i|)$, (2-sided probability), ordered so that $p_1 \leq p_2$, where t_i is the observed value of T_i, and where $\pi_i = P\,(i \text{ hypotheses rejected})\,,\ i = 0,\ 1,\ 2$.

Let $\beta = 1 - (1 - \alpha)^{(0.5)}$

Let $\gamma = (\alpha - 3\alpha^2)/(2 - 4\alpha)$

Single-step test

Reject H_i if $p_i \leq 2\beta, i = 1, 2$, and decide θ_i is sign of t_i.

Step-up test

Step 1: Reject H_1 and H_2 if $p_2 \leq 2\alpha$ and decide θ_i is sign of $t_i, i = 1, 2$;
otherwise accept H_2

Step 2: Reject H_1 if $p_1 \leq 2\gamma$ and decide θ_1 is sign of t_1

Single-step test	Step-up test
$\pi_0 = (1 - 2\beta)^2$	$\pi_0 = 1 - 2\alpha + 2\alpha^2$
$\pi_1 = \beta(1 - 2\beta)$	$\pi_1 = (\alpha - 3\alpha^2)/2$
$\pi_2 = \beta^2$	$\pi_2 = \alpha^2$

3.3.2. *$k > 2$ and/or correlated test statistics.* The methods used in Bohrer *et al* and Hochberg *et al* insure that the probability of one or more errors is $\leq \alpha$ as all values of θ approach zero. The results described in Section 3.3.1 for independent random variables, $k = 2$, carry over to correlated random variables for $k = 2$, and Hochberg *et al* note that with independent test statistics, a step-up procedure for $k = 3$ can be shown to be locally optimal under symmetry and monotonicity requirements, as for $k = 2$. However, although for $k = 2$ and independent test statistics, the $FWER$ can be shown to be $\leq \alpha$ for all values of the θ_i, Bohrer *et al* show that the $FWER$ can be $> \alpha$ for some values of θ_l, θ_2 with correlated test statistics. Furthermore, Hochberg *et al* note that for $k = 3$, even for independent test statistics, "there is difficulty in establishing the required control of the probability of any error under all θ." Thus, $FWER$ is not controlled at level α for $k = 2$, correlated test statistics, and is not known to be controlled at that level for the defined step-up test procedures for $k = 3$.

3.3.3. *More general optimality criteria.* Other regions are optimal with other optimality criteria, although some are unappealing on intuitive grounds (such as never permitting a directional decision for more than one parameter, or only permitting directional decisions jointly for both parameters.) The variety of results for different choices of local $FWER$ and different optimality criteria is too extensive to discuss in detail here. However, a generalized set of optimality criteria discussed by Hochberg *et al* in the case of two independent test statistics with $FWER$ control at $\alpha < 1/3$ connects in an interesting fashion to the one-sided results.

In order to discuss this generalization, more detail on the methods used in both Bohrer *et al* and Hochberg *et al* are necessary. In both papers, as noted previously, the procedures are discussed in terms of decision regions. As pointed out in the discussion of Lehmann (1957b) above, there are nine possible decision regions; all combinations of the three decisions for each θ_i– that it is positive (perhaps including zero), negative, or indeterminate. These regions will be labelled (a1,a2), where a_i is +, -, or 0 according to whether θ_i is called positive, negative, or indeterminate. The symmetry assumption requires that as $\theta_i \rightarrow 0$, the decision regions (0,+), (0,-), (+,0), and (-,0) approach equal probabilities, each denoted π_1, and (+,+), (+,-), (-,+) and (-,-) have equal probabilities, each denoted π_2. Without loss of generality, assume θ_1, θ_2 are positive. The probability of the decision region (0,0) is denoted π_0. Given the symmetry assumptions, local control of directional error requires $2\pi_1 + 3\pi_2 \leq \alpha$, and probability theory requires $\pi_0 + 4\pi_1 + 4\pi_2 = 1$.

The expected number of correct sign designations is $2\pi_1 + 4\pi_2$. Hochberg and Posner consider a generalization, $a\pi_1 + b\pi_2$, giving arbitrary weights to decisions with one correct sign designation and two correct sign designations. Given the symmetry and probability requirements, the solution requires maximizing $a\pi_1 + b\pi_2$ subject to

(i) $2\pi_1 + 3\pi_2 \leq \alpha$

(ii) $\pi_2 \leq \alpha^2$

(iii) $\pi_1 + \pi_2 \leq 1/4$

(iv) $\pi_1, \pi_2 \geq 0$.

They obtain results for $a/b < 2/3$ and for $a/b > 2/3$. When $a/b < 2/3$, the optimality results are the same as for maximizing the number of correct sign designations; a two-sided step-up procedure is optimal as in Table 2.

When $a/b > 2/3$, and $\alpha < 1/2$, a single-step procedure is optimal. This criterion $(a/b > 2/3)$ seems undesirable, since it favors decisions with the sign of one parameter designated over those with the sign of both parameters designated when the $FWER$ restriction permits both to be designated.

How about $a/b = 2/3$? Hochberg and Posner state in the last paragraph:

"We have omitted the case of $a/b = 2/3$ from our discussion. When $\alpha < 3/4$, the optimal solution to this case will not be unique. Thus, examination of additional criteria will be required to fix π_1 and π_2. We leave this to future research."

Note that $a/b = 2/3$ is the criterion of maximizing the local probability of at least one correct sign designation. Note that it is also the local probability of at least one incorrect sign designation, the $FWER$ at the origin, which is also the global $FWER$ in the independent case with $k = 2$. Without loss of generality we can take $a = 2$ and $b = 3$, in which case the maximum by requirement (i) is α, and the maximum is attained by any procedure that controls the maximum $FWER$ exactly at α. Given all such procedures, note that the expected number of correct sign designations equals $\alpha + \pi_2$. Therefore, a solution that maximizes π_2 at α^2 (requirement (iii)) is optimal both for maximizing the probability of at least one correct sign and for maximizing the expected number of correct signs, another way to look at the optimality property of solutions in this class, including the optimal step-up procedure in Table 2.

3.3.4. *Further considerations, $FWER$ control.* When only directional errors are considered, the two-sided step-down procedure is not optimal under any of the optimality criteria considered. Bohrer and Schervish note that under their optimality criterion, maximizing the number of correct sign designations, the two-sided step-down procedure is optimal if the (0,0) region is required to be square, but that requirement is not especially appealing. However, if nondirectional errors are the only ones considered, neither the two-sided step-down nor step-up procedure dominates the other (Dunnett and Tamhane, 1992), so it is possible that step-down procedures may be optimal under some reasonable criteria in the two-sided case under formulation D1.

As is the case with the Lehmann (1952) paper, Bohrer *et al* and Hochberg *et al* described the procedures in terms of multiple decision regions of the sample space. The alternative way of thinking about such procedures and describing them as step-down or step-up hypothesis testing procedures began much later. Both points of view can be useful in giving insight into properties of procedures.

3.4. **False discovery rate control.** Recently, a new criterion has been proposed for error control in multiple testing: the false discovery rate (Benjamini and Hochberg, 1995). This is defined as the expected value of $Q = V/R$, where V is the number of true hypotheses that are rejected (i.e. Type I errors), and R is the total number of rejected hypotheses. When $R = 0$, Q is defined as 0. Benjamini and Hochberg (1995) proposed a step-up method that controls the FDR at a designated level α for independent test statistics, and has been shown to control it for some types of dependent test statistics. For two-sided tests, these results hold under the two-sided formulation D1.

When all hypotheses are true, $FWER$ and FDR procedures using the same level of α both control the $FWER$ at that level. However, when some hypotheses are false, the probability of rejecting hypotheses is increased using the more lenient criterion of FDR as opposed to $FWER$ control. This alternative FDR-controlling criterion seems especially reasonable when massive numbers of hypotheses are being tested and when many are assumed false, such as in microarray analysis in genomics, and when positive results will be followed up by further investigation before final acceptance.

There is a great deal of current work on properties of the false discovery rate, including some on optimality properties. Benjamini and Hochberg (1995) described a post-hoc choice of α as satisfying a constrained optimization problem. Abramovich, Benjamini, Donoho, and Johnstone (2000) make some connections between the use of an FDR method, asymptotic minimaxity, model selection and decision-theoretic analysis in wavelet image analysis. In a Bayesian context Genovese and Wasserman (2002), Efron, Storey, and Tibshirani (2001), and Storey (2002) have some optimality results. This area is changing fast, and a description is beyond the scope of this paper.

4. Bayesian approaches

Duncan, in a series of papers beginning in the 1960's (see, e.g., Duncan, 1961, Waller and Duncan, 1969, Duncan and Dixon, 1983)), developed a Bayesian procedure for comparing means of normal distributions based on the assumption that the true means are realizations of independent normal random variables. He used Lehmann's (1957b) theorem with a somewhat different loss structure to show that his procedure has minimum Bayes risk.

The modified loss functions are as follows: Instead of loss a for incorrect rejection, the loss is $a|\theta_1 - \theta_2|$, and instead of loss b for incorrect acceptance, the loss is $b|\theta_1 - \theta_2|$. In other words, the loss depends on the magnitude of the incorrect decision–the loss in saying $\theta_1 - \theta_2 < 0$ when the true difference is positive and large is greater than when it is positive and small.

Lewis (1997) adopted the Bayesian assumption of Duncan on the distribution of true means, but controls the maximum $PFER$ averaged over the distribution of means. Shaffer (1999) adapted the Duncan approach to require the procedure to control the maximum $FWER$. Shaffer found surprising similarity in power functions between this Duncan adaptation and the FDR-controlling test of Benjamini and Hochberg (1995). Lewis and Thayer (2002) give some heuristic arguments to account for this relationship.

Since in Duncan's formulation the loss when testing H_i is zero if $\theta_i = 0$, regardless of the outcome, directional errors are the only ones that are counted as losses. This is one of two main approaches Bayesians have taken. Recently, Gelman and Tuerlinckx (2002) and Lewis and Thayer (2002) have considered the same structure as Duncan, but used 0-1 loss functions for directional errors, as in Lehmann's approach, without considering the magnitude of those errors. The Bayesian solution, supplemented by corrections for multiplicity, is very much like that in Lehmann's frequentist approach, but has an extra term depending on the prior variance of the true mean distribution. The Duncan, Lewis, Shaffer, Gelman and Tuerlinckx, and Lewis and Thayer approaches are similar in that under all of them the prior probabilities that the θ_i's $= 0$ are zero, so that only directional (Type III) errors are possible.

Berger and Sellke (1987) used a different approach that allows for the possibility $\theta = 0$ by introducing a prior with a point mass probability at $\theta = 0$. They obtain the posterior probability that $\theta = 0$ under a variety of assumptions on the remainder of the prior distribution. This leads to interesting comparisons with classical p-values, beyond the scope of this paper.

5. Conclusions and Summary

Optimality in multiple testing has been considered under a variety of criteria and conditions. Results for this problem have been obtained using different generalizations of the Neyman-Pearson single-hypothesis concepts of Type I error and power, from the point of view of multiple decision theory, from Bayesian positions, and from mixtures of these.

This paper restricts attention to testing hypotheses related to the values of k parameters, $k > 1$, k finite, and relates current practice to some early literature.

There are three principal conditions that affect the types of procedures that have been shown to be optimal for this class of hypotheses.

1. Type of error control

The paper has restricted attention primarily to two types of error control: $PFER$ control and $FWER$ control. The recent introduction of another criterion, FDR control, has led to a series of new papers, some of which deal with optimality under that criterion. That area is developing rapidly now. The definition and some references are given, but otherwise FDR control is not covered in the present paper.

Under the additive criterion of $PFER$ control, single-stage procedures have been found to be optimal under a variety of optimality criteria for both one-sided and two-sided tests. and in the two-sided case, for procedures controlling both directional (Type III) and nondirectional (Type I) errors. The results appear to hold for arbitrary k and many hold under conditions of dependence as well as independence.

Control of $FWER$, a non-additive criterion, is more difficult to treat statistically. Under $FWER$ control, stepwise procedures dominate single-step procedures. There are fewer results under this criterion, and they depend on other considerations, as noted below.

2. $FWER$-control: One-sided versus two-sided procedures.

Symmetry and monotonicity (upper convexity) in various forms are common assumptions in seeking optimal procedures. For one-sided procedures, Lehmann (1952) proved that under relatively general conditions, for $k = 2$, a step-down procedure is optimal in minimizing the maximum probabilities of errors of various kinds, given various priorities on decision regions. Some work is now being pursued that suggests that if the priorities are changed, a step-up procedure is optimal. It appears from currently ongoing research that these results can be extended to $k \geq 2$, and they do not require independence.

3. $FWER$-control: Two-sided hypotheses, nondirectional and directional errors.

Some approaches to the two-sided test assume no null values, or at least no additional losses under null values. Then directional errors only are considered. This makes the error criterion continuous as $\theta_i \rightarrow 0$ for all i. Under this assumption, step-up test procedures have been found to be optimal in locally maximizing the expected number of correct sign classifications for independent random variables, $k = 2$. In contrast to the one-sided case, the results for $k > 2$ and for dependent random variables are more problematic, since local control of the $FWER$ at α does not imply global control at that value in these generalized situations.

Thus, optimality results for $FWER$ control of Type III errors in testing two-sided hypotheses are very limited, and many open questions remain. Optimality results for control of Type I (nondirectional) errors, and for joint control of Type I and Type III (directional) errors, appear to be open problems.

References

[1] Abramovich, F., Benjamini, Y., Donoho, D., and Johnstone, I. (2000). Adapting to unknown sparsity by controlling the false discovery rate. Technical Report, Department of Statistics, Stanford University.

[2] Beaton, A.E. and Zwick, R. (1992). Overview of the National Assessment of Educational Progress. *Journal of Educational Statistics, 17*, 95-109.

[3] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate. *Journal of the Royal Statistical Society, 57*, 289-300.

[4] Berger J. and Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of P-values and evidence (with discussion). *Journal of the American Statistical Association 82*, 112-139.

[5] Berger R. (1989). Uniformly more powerful tests for hypotheses concerning linear inequalities and normal means. *Journal of the American Statistical Association 84*, 192-199.

[6] Bohrer, R. (1979). Multiple three-decision rules for parametric signs. *Journal of the American Statistical Association, 74*, 432-437.

[7] Bohrer, R. (1982). Optimal multiple decision problems: Some principles and procedures applicable in cancer drug screening. In L. LeCam, & J. Neyman (Eds.), *Probability Models for Cancer.* Amsterdam:North Holland, 287-301.

[8] Bohrer, R., and Schervish, M. (1980). An optimal multiple decision rule about signs. *Proceedings of the National Academy of Sciences, 77*, 52-56.

[9] Duncan, D.B. (1961). Bayes rules for a common multiple comparisons problem and related Student-t problems. *Annals of Mathematical Statistics 32*, 1013-1033.

[10] Duncan, D.B. (1965). A Bayesian approach to multiple comparisons. *Technometrics, 7*, 171-222.

[11] Duncan, D.B. and Dixon, D.O. (1983). k-ratio t tests, t intervals, and point estimates for multiple comparisons. *Encyclopedia of Statistical Sciences, 4*, ed. S. Kotz, N.L. Johnson, 403-410.

[12] Dunnett, C. W., and Tamhane, A. C. (1992). A step-up multiple test procedure. *Journal of the American Statistical Association, 87*, 162-170.

[13] Efron, B., Storey, J.D., and Tibshirani, R. (2001). Microarrays, empirical Bayes methods, and false discovery rates. Technical Report 2001-218, Department of Statistics, Stanford University.

[14] Finner, H. (1999). Stepwise multiple test procedures and control of directional errors. *Annals of Statistics, 27*, 274-289.

[15] Gelman, A. and Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics, 15*, 373-390.

[16] Genovese, C., and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society Series B–Statistical Methodology, 64*, 499-517.

[17] Hochberg, Y. (1987). Multiple classification rules for signs of parameters. *Journal of Statistical Planning and Inference, 15*, 177-188.

[18] Hochberg, Y. and Posner, M.E. (1986). On optimal decision rules for signs of parameters. *Annals of Statistics, 14*, 733-742.

[19] Hochberg, Y., and Tamhane, A.C. (1987). *Multiple Comparison Procedures*, New York: Wiley.

[20] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65-70.

[21] Jones, L. V., and Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods, 5*, 411-414.

[22] Klockars, A. J., and Hancock, G. R. (1994). Per-experiment error rates: The hidden costs of several multiple comparison procedures. *Educational and Psychological Measurement, 54*, 292-298.

[23] Lehmann, E. L. (1951). A general concept of unbiasedness. *Annals of Mathematical Statistics 22*, 587-592.

[24] Lehmann, E. L. (1952). Testing multiparameter hypotheses. *Annals of Mathematical Statistics 23*, 541-552.

[25] Lehmann, E.L. (1957a). A theory of some multiple decision problems Part I. *Annals of Mathematical Statistics, 28*, 1-25.

[26] Lehmann, E.L. (1957b). A theory of some multiple decision problems Part II. *Annals of Mathematical Statistics, 28*, 547-572.

[27] Lehmann, E.L. & Shaffer, J.P. (1979). Optimum significance levels for multistage comparison procedures. *Annals of Statistics, 7*, 27-45.

[28] Lewis, C. (1997) Qualitative Multiple comparisons: Fisher revisited. Unpublished manuscript.

[29] Lewis, C. & Thayer, D. (2002). Multiple inferences for random effects. Paper presented at the Third International Conference on Multiple Comparisons, Bethesda, MD.

[30] Liu, W. (1997). Control of directional errors with step-up multiple tests. *Statistics & Probability Letters, 31*, 239- 242.

[31] Mosteller, F. (1948). A k-sample slippage test for an extreme population. *Annals of Mathematical Statistics, 19*, 58-65.

[32] Paulson, E. (1952). An optimum solution to the k-sample slippage problem for the normal distribution. *Annals of Mathematical Statistics 23*, 610-616.

[33] Perlman, M.D. and Wu, L. (1999). The emperor's new tests: A defense of the likelihood ratio criterion (with discussion). *Statistical Science 14*, 355- 381.

[34] Rom, D.M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometria, 77*, 663-665.

[35] Ryan, T.A. (1959). Multiple comparisons in psychological research. *Psychological Bulletin, 56*, 26-47.

[36] Ryan, T.A. (1960). Significance tests for multiple comparison of proportions, variances, and other statistics. *Psychological Bulletin, 57*, 318-328.

[37] Scheffé, H. (1970). Multiple testing versus multiple estimation. Improper confidence sets. Estimation of directions and ratios. *Annals of Mathematical Statistics, 41*, 1-29.

[38] Shaffer, J.P. (1980). Control of directional errors with stagewise multiple test procedures. *Annals of Statistics, 8*, 1342-1347.

[39] Shaffer, J.P. (1999). A semi-Bayesian study of Duncan's Bayesian multiple comparison procedure. *Journal of Statistical Planning and Inference, 82*, 197-213.

[40] Shaffer, J.P. (2002). Multiplicity, directional (Type III) errors, and the null hypothesis. *Psychological Methods 7*, 356-369.

[41] Schwager, S.J. (1985). Mean slippage problems. *Encyclopedia of Statistical Sciences, 5*, 372-375.

[42] Spjøtvoll, E. (1972). On the optimality of some multiple comparison procedures. *Annals of Mathematical Statistics, 43*, 398-411.

[43] Storey, J.D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 64*, 479-498.

[44] Tamhane, A.C., Liu, W., and Dunnett, C.W. (1998). A generalized step-up-down multiple test procedure. *Canadian Journal of Statistics, 26*, 353-363.

[45] Truax, D. (1953). An optimum slippage test for the variances of k normal distributions. *Annals of Mathematical Statistics, 24*, 669-674.

[46] Tukey, J.W. (1952). Reminder sheets for "Multiple Comparisons". In Braun, H.I. (Ed.), *The Collected Works of John W. Tukey, Vol. VIII*, New York: Chapman & Hall, 341-345.

[47] Tukey, J.W. (1953). The problem of multiple comparisons. In Braun, H.I. (Ed.), *The Collected Works of John W. Tukey, Vol. VIII*, New York: Chapman & Hall, 1-300.

[48] Waller, R.A., and Duncan, D. B. (1969). A Bayes rule for the symmetric multiple comparisons problem. *Journal of the American Statistical Association, 64*, 1484-1503.

[49] Zelterman, D. (1990) On tests for qualitative interactions. *Statistics & Probability Letters 10*, 59-63.

Juliet Popper Shaffer
Department of Statistics
367 Evans Hall
University of California
Berkeley, CA 94720

shaffer@stat.berkeley.edu