Erich Lehmann's contributions to multiple decision making

Juliet Popper Shaffer

1 Introduction

Much of statistical theory and methodology deals with testing a single hypothesis or specifying a single confidence set. But many, probably most studies deal with a number of questions, requiring more than one test or confidence set. In fact, within the last 20 years, advances in scientific and computer technology has made it possible to collect and analyze massive amounts of data, resulting in decision-making for thousands, even millions, of items.

Should each hypothesis or confidence set be treated individually, without taking the number of such inferences into account? Under what conditions would that be satisfactory, and when would it be inadvisable? If inadvisable, what alternative approach should be taken? These considerations constitute the subject of multiple inference, to which Erich Lehmann has made many contributions.

There was sporadic work in this area earlier, but concentrated efforts to deal with multiplicity issues in a comprehensive way began around 1950, with papers by Tukey, Duncan, Scheffé, and others. See, e.g. Hochberg and Tamhane (1987) for descriptions. After a brief discussion in his 1952 paper "Testing multiparameter hypotheses," Erich's detailed work on this problem began with his 1957 papers.

In these comments, a consistent notation will be used, rather than changing notation as it appears in each paper.

J.P. Shaffer

Department of Statistics, University of California, Berkeley, CA 94720-3860
e-mail: shaffer@stat.berkeley.edu

2 The 1957 papers

In the first of two papers, “A theory of multiple decision problems” I and II in the *Annals of Statistics* (Lehmann, 1957a), Erich begins with an example of two independent, normally-distributed random variables X_1 and X_2 with unit variance, where $E(X_i) = \theta_i$, $i = 1, 2$. If we were interested in testing $H : \theta_1 = 0$ and $\theta_2 = 0$ versus ($\bar{H}$): at least one $\theta_i \neq 0$, it would seem sensible to use the spherically symmetric rejection region $x_1^2 + x_2^2 > c$. However, if upon rejection more detailed statements were desired, we would have a multiple decision problem. The parameter space can be partitioned into the nine regions corresponding to all combinations of the decisions for each θ_i : that it equals 0, is less than 0, or is greater than 0. Then a multiple decision procedure should permit partitioning the regions of the parameter space to allow the possibility of each of these decisions. These can be generated by testing, for each i , the hypotheses $H_{i1} : \theta_i \leq 0$ and $H_{i2} : \theta_i \geq 0$, with acceptance of both H_{i1} and H_{i2} equivalent to the decision $\theta_i = 0$.

The proposed loss structure is as follows:

Loss = 0 for a correct decision, a for rejecting a true hypothesis, b for accepting a false hypothesis. The loss table for the set of decisions is obtained by adding the losses for the individual decisions. If each hypothesis H_{i1} and H_{i2} is tested at level $b/(a+b)$ using the standard t -test, the natural resulting multiple test procedure is shown to be unbiased and to have uniformly minimum risk.

The procedure is described in much more general terms, given any composite hypothesis that, if rejected, can be partitioned into a set of detailed hypotheses in this way. The only requirements are that each separate hypothesis be tested by a method with uniformly minimum risk, and that the set of separate decisions is *compatible*, i.e., that there are no sets in the decision space with positive probability corresponding to sets in the parameter space with zero probability, and vice versa.

The formulation is shown to apply to a large number of problems, including point estimation, goodness of fit, and equivalence (rather than equality) testing. Thus, a good deal of literature in general statistics is encompassed.

In the second of these papers (Lehmann, 1957b), Erich reformulates the problem, noting that the earlier formulation does not generalize to a number of areas including the very important area of comparison of pairs of parameters, in which the compatibility assumption is not satisfied.

For example, suppose we have random samples from three normal populations and want to test the three pairs of hypotheses

- (i) $\theta_1 - \theta_2 \leq (\geq) 0$,
- (ii) $\theta_1 - \theta_3 \leq (\geq) 0$,
- (iii) $\theta_2 - \theta_3 \leq (\geq) 0$.

If the uniformly-minimum-risk tests of each of the six hypotheses are used, it is possible to arrive at the decision

$$\theta_1 < \theta_3, \theta_1 = \theta_2, \theta_2 = \theta_3, \quad (2.1)$$

while there is no set in the parameter space corresponding to this decision. Thus, this set of decisions is not compatible.

In the second 1957 paper, the decision to accept a hypothesis is interpreted, not as an assertion that the hypothesis is true, but as providing no information about the value of the parameter involved. Instead of the conclusion (1) in this example, a decision with positive probability although there is no corresponding state of nature, the conclusion would be

$$-\infty < \theta_1 - \theta_2 < \infty, -\infty < \theta_2 - \theta_3 < \infty, \theta_1 < \theta_3. \quad (2.2)$$

The reformulation includes the possibility of such decisions, treated as partial joint decisions, and the earlier optimality properties still hold.

Erich acknowledges early formulations by Duncan that are similar, but not identical, to this approach. (Duncan's loss functions consider the magnitude as well as the direction of departure from the null hypothesis.) In return, Duncan, in a series of papers beginning in the 1960s (see Duncan, 1961), developed a Bayesian procedure for comparing means of normal distributions based on the assumption that the population means are realizations of independent normally-distributed random variables. He used Erich's theorem to show that his procedure has minimum Bayes risk. Shaffer (1999) studied Duncan's work in an empirical Bayes and frequentist context. Pennello (2007) has provided an overview of Duncan's k -ratio Bayes rule approach and further developments arising out of it.

3 Optimum significance levels for multistage comparison procedures

Much of the work in multiple testing has used not the additive risk formulation of the 1957 papers, but rather has been based on the idea first proposed by Tukey of controlling the familywise error rate ($FWER$), *i.e.*, the probability of one or more errors in the whole set of comparisons, at some small value α , typically 0.05. Therefore the level at which any hypothesis is tested may depend on the number of and relationships among tested hypotheses, rather than only the individual hypothesis loss structure. In the so-called stepwise methods, the level for testing any hypothesis also depends on the outcomes of tests of other hypotheses. This paper (Lehmann and Shaffer, 1979) deals with multiple range procedures, one type of stepwise method applicable to comparisons of the effects of treatments. Some optimality conditions are proposed and a procedure derived that turns out to be almost identical to a widely-used procedure proposed by Tukey, who did not note any optimality properties. The multiple range methods begin by testing the equality of the full set of treatments, comparing the treatments with the largest and smallest sample means using the upper- α significance level of the range or studentized range, proceeding to successively smaller ranges only if the including ranges have been declared significant. The paper finds the required significance levels for each successive range test, with the criterion of maximizing the levels for the smallest ranges. An alternative optimality criterion, maximizing the successive levels for the largest ranges, was investigated by Finner (1990).

Many later developments in stepwise testing of pairwise comparisons with *FWER* control have been discussed in textbooks (Hochberg and Tamhane, 1987, Westfall and Young, 1993, Hsu, 1996) and in articles (Shaffer, 1995, Tamhane, 1996).

4 On optimality of stepdown and stepup multiple test procedures

4.1 Previous related work

In the 1952 paper mentioned in Section 1, “Testing multiparameter hypotheses”, instead of considering hypotheses $H_i : \theta_i \leq \theta_{i0}, i = 1, \dots, s$, with the desire to find all false H_i , Erich considers the hypothesis H_1 or H_2 or $\dots H_s$ are true, with the alternative that all H_i are false. If $\theta_{i0} = 0$ for all i , the alternative is equivalent to stating that $\min(\theta_i) > 0$; *i.e.*, that all θ_i are positive. This kind of test also has important applications. For example, in medicine it is vital to know whether a drug is helpful not only on the average but also for subgroups of a population, divided by age, gender, ethnicity, etc., *i.e.*, whether the effects are positive for all designated groups.

For simplicity, assume two groups, and assume X_1 and X_2 are independently distributed $N(\theta_i, 1)$, $i = 1, 2$, respectively. In considering this type of hypothesis, Erich defines a *monotone* procedure as one for which, if x_1 and x_2 are in the rejection region Ω and $x'_1 \geq x_1, x'_2 \geq x_2$, then x'_1 and x'_2 are in Ω . This seems desirable if large values of the test statistics are more likely with large values of the parameters, as in this example. Then Erich proves that the test procedure: Reject H if $\min(x_i) > u_\alpha$, where u_α is the upper- α critical value of the normal distribution, is the uniformly most powerful monotone test with $FWER = \alpha$ for the hypothesis.

Note that the optimal monotone test is extremely conservative for small values of the parameters: Power has a lower limit of α^2 if the test statistics are independent, as θ_1 and θ_2 approach 0 from above, and a lower limit of α^s for s such hypotheses. He considers the monotonicity requirement to be sensible, but points out that more powerful procedures are possible when this requirement is dropped, and describes one such procedure (p. 542-43), which was adopted by Roger Berger (1989) and generalized in a series of papers. The idea is to reject when *both* test statistics are within certain designated small regions. There is related work by Gutmann (1987) and Zelterman (1990). Perlman and Wu (1999) has additional references and alleges that such procedures, although uniformly more powerful than monotone procedures, defy common sense and should not be used.

4.2 Application to this paper

Holm (1979) introduced a general stepdown procedure with *FWER* control that can be used with any set of hypotheses, not only pairwise equality comparisons. It is a modification of the well-known Bonferroni method, sometimes called the simultaneous or singlestep method, in which each of s hypotheses is rejected at level α if its associated p -value is $\leq \alpha/s$. Assume the hypotheses are numbered in the order of their p -values, i.e., H_i is the hypothesis indexed by p_i , $p_1 \leq p_2 \leq \dots \leq p_s$. In Holm's stepdown procedure at level α , H_i is rejected if and only if $p_i < \alpha/(s-i+1)$, $\forall i \leq s$. The procedure can be carried out by beginning with H_1 and successively rejecting the ordered hypotheses as long as each successive hypothesis is rejected.

With this procedure, *FWER* is $\leq \alpha$ with no restrictive conditions. Later, a stepup procedure was introduced for normally-distributed statistics (Dunnett and Tamhane, 1992), and under more general conditions by Hochberg (1988), the latter based on work by Simes (1986). There is a great deal of literature dealing with these methods, as well as a further extension to generalized step-up-down methods, introduced by Tamhane, Liu, and Dunnett (1998). In contrast with the Holm stepdown method, the stepup and generalized step-up-down methods require some restrictions on the joint distributions of the test statistics for *FWER* control. See Sarkar (2007, 2008) for a review of these methods and required conditions.

Lehmann, Romano, and Shaffer (2005) described and proved the first optimality results for *FWER*-controlling stepup and stepdown methods based on individual p -values. Under monotonicity conditions including that noted above in the discussion of the 1952 paper, and considering one-sided hypotheses, it is shown that the procedures are optimal in the sense that stepdown procedures maximize the minimum probability of rejecting at least one false hypothesis, and stepup procedures maximize the minimum probability of rejecting all false hypotheses. As of this writing, a similar minimax result has not been extended to the step-up-down procedures, but it seems a reasonable possibility that it could be.

5 Generalizations of the Family-wise Error Rate

Although tests of large numbers of hypotheses had been common in areas such as survey sample analyses and epidemiological studies, the somewhat complex structure of these studies had led to either ignoring the multiplicity issues or treating them in an ad hoc way, dividing up the tests into families and using *FWER* control within those families. The development within the last 20 years of technology for mass observations in fields like genomics, neuroscience, and astronomy has led to the desire for joint tests of thousands, even millions, of hypotheses as a single family. The problem of error buildup in these cases, both rejection of true hypotheses and acceptance of false ones, was obvious and couldn't be ignored. Furthermore, for most of these applications, the consequences of rejecting true hypotheses were likely to be less serious than in

previous multiple hypothesis tests, since apparent discoveries would result in followup studies before any action would be taken. Familywise error control, by assuring very low probability of even a single error of rejecting a true hypothesis, results in a correspondingly low probability of rejecting false hypotheses, *i.e.*, of power.

For these new applications, the *FWER*-controlling methods seemed too conservative, but the need for some multiple error control was clear. As a result, interest in multiple testing blossomed, and a series of alternative criteria have been developed and explored.

Although there were earlier explorations of alternative criteria (Victor, 1982, Hommel and Hoffman, 1988, Seeger, 1968, Eklund and Seeger, 1965), the first paper to receive widespread attention was that of Benjamini and Hochberg (1995), who proposed use of the false discovery rate (*FDR*) (earlier treated less formally in Seeger, 1968, and Eklund and Seeger, 1965)), an alternative, more liberal, criterion than the *FWER*, and presented a method, based on an equality of Simes (1986), for controlling it.

Letting R = the number of rejections, V = the number of false rejections (*i.e.*, rejections of true hypotheses), the *FDR* is defined as $E(V/R | R > 0) * Pr(R > 0)$. This equals the *FWER* if all hypotheses are true, but allows some false rejections as long as their expected proportion is no greater than a specified, usually small, value. Many related measures have since been explored in detail, including: control of the positive *FDR* (Storey), control of $Pr(V \geq k)$, called the $k - FWER$, where k is a specified integer (note: $k - FWER = FWER$ when $k = 1$), $Pr(V/R > \gamma)$, where V/R is called the false discovery proportion (*FDP*), and γ is a specified proportion, various measures related to the probabilities of nonrejection of true hypotheses and balances between nonrejection of true and rejection of false hypotheses, Bayes and empirical Bayes approaches, and methods justified by good asymptotic properties in relation to many of these. See Farcomeni (2008), a review article with many references.

Lehmann and Romano (2005) proposed methods of controlling the number of false rejections - some of their results for this criterion were obtained earlier by Hommel and Hoffman (1988)-, and associated methods of controlling the *FDP*. The methods can be contrasted with those proposed by van der Laan, Dudoit, and Pollard (2004). Romano and Wolf (2007) compares the two types of methods, generalizations, and alternatives theoretically and empirically. See also Sarkar (2007, 2009).

Let $p_i, i = 1, \dots, s$, be the ordered p -values of the s hypotheses being tested, $p_1 \leq p_2 \leq \dots, p_s$, with the correspondingly ordered hypotheses $H_1, \dots, H_s$. Let $a_i, i = 1, \dots, s$ be constants associated with the respective p -values.

k -*FWER* control: In Method 1, $a_i = k\alpha/s, \forall i$, and all hypotheses for which $p_i < a_i$ are rejected. When $k = 1$, this is the well-known Bonferroni method of controlling the *FWER*, and is often called the simultaneous or single-step method.

In Method 2, $a_i = k\alpha/s, i \leq k$ and $a_i = k\alpha/(s + k - i), i > k$. Reject all hypotheses $H_1, \dots, H_j$, where j is the largest integer such that $p_i \leq a_i \forall i = 1, \dots, j$. When $k = 1$, it corresponds to the Holm method for *FWER* control, and thus is a generalization of the stepdown method to control of the $k - FWER$.

One somewhat unpleasant consequence of sole reliance on $k - FWER$ control is the fact that one can reject $k - 1$ hypotheses without any consideration of the associated

p -values and still meet the criterion. Even if monotonicity in the p -values is required, their specific values can be ignored. The paper discusses this possibility but notes that k or more hypotheses can be rejected if and only if $p_k \leq k\alpha/s$. At any rate, it seems desirable for some additional criteria besides $k - FWER$ control to be imposed.

As is typical in papers authored or co-authored by Erich, an optimality result is proved: No a_i can be increased without violating $k - FWER$ control for some joint distribution of the test statistics.

Romano has since explored more powerful modifications (Romano and Shaikh, 2006a, 2006b), Romano and Wolf (2007). As noted above, the latter article includes comparisons of a number of $k - FWER$ and FDP -controlling methods.

6 Summary

Erich worked early in the area of multiple decision making. As shown in the discussions of the individual papers, his work has been influential in many facets of this topic. His papers have given results for types of problems of practical importance that are the subject of ongoing research. His work in this area always involves optimality considerations. One expression that can never be used in relation to Erich's work is: *ad hoc*.

References

- [1] Benjamini, Y. and Hochberg, Y. (1995). "Controlling the false discovery rate." *Journal of the Royal Statistical Society Series B*, 57, 289-300.
- [2] Berger, R. L. (1989). "Uniformly more powerful tests for hypotheses concerning linear inequalities and normal means." *Journal of the American Statistical Association*, 84, 192-199.
- [3] Duncan, D. B. (1961). "Bayes rules for a common multiple comparison problem and related Student-t problems." *Annals of Mathematical Statistics*, 32, 1013-1033.
- [4] Dunnett, C. W. and Tamhane, A. C. (1992). "A step-up multiple test procedure." *Journal of the American Statistical Association*, 87, 162-170.
- [5] Eklund, G. and Seeger, P. (1965). Masssignifikansanalys. *Statistisk Tidskrift*, 3rd series-4, 355-365.
- [6] Farcomeni, A. (2008). "A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion." *Statistical Methods in Medical Research*, 17, 347-388.
- [7] Finner, H. (1990). "Some new inequalities for the range distribution, with application to the determination of optimum significance levels of multiple range tests." *J. Amer. Statist. Assoc.*, 85, 191-194.
- [8] Gutmann, S. (1987). "Tests uniformly more powerful than uniformly most powerful monotone tests." *Journal of Statistical Planning and Inference*, 17, 279-292.
- [9] Hochberg, Y. (1988). "A sharper Bonferroni procedure for multiple tests of significance." *Biometrika*, 75, 800-802.
- [10] Hochberg, Y. and Tamhane, A. (1987). "Multiple Comparison Procedures." Wiley, New York.
- [11] Holm, S. (1979). "A simple sequentially rejective multiple test procedure". *Scandinavian Journal of Statistics*, 6, 65-70.

- [12] Hommel, G. and Hoffman, T. (1988). "Controlled uncertainty." In *Multiple Hypotheses Testing* (P. Bauer, G. Hommel, and E. Sonneman, Eds.), 154-161. Springer, Heidelberg.
- [13] Hsu, J. C. (1996). *Multiple Comparisons: Theory and Methods*. Chapman & Hall, New York.
- [14] Lehmann, E. L. (1952). "Testing multiparameter hypotheses". *Annals of Mathematical Statistics*, 23, 541-552
- [15] Lehmann, E. L. (1957a). "A theory of some multiple decision problems I". *Annals of Mathematical Statistics*, 28, 1-25.
- [16] Lehmann, E. L. (1957b). "A theory of some multiple decisions problems II". *Annals of Mathematical Statistics*, 28, 547-572.
- [17] Lehmann, E. L. and Shaffer, J. P. (1979). "Optimum significance levels for multistage comparison procedures". *Annals of Statistics*, 7, 27-45.
- [18] Lehmann, E. L. and Romano, J. P. (2005). "Generalizations of the familywise error rate". *Annals of Statistics*, 33, 1138-1154.
- [19] Lehmann, E. L., Romano, J. P., and Shaffer, J. P. (2005). "On optimality of stepdown and stepup multiple test procedures". *Annals of Statistics* 33, 1138-1154.
- [20] Pennello, G. A. (2007). "Duncan's k-ratio Bayes rule approach to multiple comparisons: An overview." *Biometrical Journal*, 49, 78-93.
- [21] Perlman, M. D. and Wu, L. (1999). "The emperor's new tests" (with discussion). *Statistical Science*, 14, 355-381.
- [22] Romano, J. P. and Shaikh, A. M. (2006a). "On stepdown control of the false discovery proportion." In *Optimality: The Second Erich L. Lehmann Symposium*, (J. Rojo, Ed), 33-50.
- [23] Romano, J. P. and Shaikh, A. M. (2006b). "Stepup procedures for control of generalizations of the family error rate." *Annals of Statistics*, 34, 1850-1873.
- [24] Romano, J. P. and Wolf, M. (2007). "Control of generalized error rates in multiple testing." *Annals of Statistics*, 35, 1378-1408.
- [25] Sarkar, S. K. (2007). "Stepup procedures controlling generalized FWER and generalized FDR." *Annals of Statistics*, 2007, 35, 2405-2420.
- [26] Sarkar, S. K. (2008). "On the Simes inequality and its genralization." In *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen*, IMS Collections Vol. 1, 231-242.
- [27] Sarkar, S. K. (2009). "On Methods Controlling the False Discovery Rate (with discussion)." *Sankhya, Ser A*. In Press.
- [28] Seeger, P. (1968). "A note on a method for the analysis of significances en masse." *Technometrics*, 10, 586-593.
- [29] Shaffer, J. P. (1995). "Multiple hypothesis testing." In *Annual Review of Psychology*, 46, 561-584.
- [30] Shaffer, J. P. (1999). "A semi-Bayesian study of Duncan's Bayesian multiple comparison procedure." *Journal of Statistical Planning and Inference*, 82, 197-213.
- [31] Simes, R. (1986). "An improved Bonferroni procedure for multiple tests of significance." *Biometrika*, 73, 751-754.
- [32] Tamhane, A. C. (1996). "Multiple Comparisons." In *Handbook of Statistics*, vol. 13 (Eds. S. Ghosh and C. R. Rao), 587-630.
- [33] 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.
- [34] van der Laan, M., Dudoit, S., and Pollard, K. (2004). "Multiple testing, Part III, Procedures for control of the generalized familywise error rate and proportion of false positives.
- [35] Victor, N. (1982). "Exploratory data analysis and clinical research." *Methods of Information in Medicine*, 21, 53-54.
- [36] Westfall, P. H. and Young, S. S. (1993). *Resampling-based Multiple Testing: Examples and Methods for p-value Adjustment*. Wiley, New York.
- [37] Zelterman, D. (1990). "On tests for qualitative interactions." *Statistics and Probability Letters*, 10, 59-63.