Journal of Modern Applied Statistical Methods

Volume 15 | Issue 1 Article 2

5-2016

The Composite Hypothesis Contrast Procedure: A Novel Sequential Multiple-Comparison Approach

Joel R. Levin
University of Arizona, jrlevin@u.arizona.edu

Ronald C. Serlin
University of Wisconsin-Madison, rcserlin@wisc.edu

Recommended Citation
Levin, Joel R. and Serlin, Ronald C. (2016) "The Composite Hypothesis Contrast Procedure: A Novel Sequential Multiple-Comparison Approach," *Journal of Modern Applied Statistical Methods*: Vol. 15 : Iss. 1 , Article 2.
DOI: 10.22237/jmasm/1462075260

The Composite Hypothesis Contrast Procedure: A Novel Sequential Multiple-Comparison Approach

Cover Page Footnote
Correspondence concerning this article should be addressed to Joel R. Levin at jrlevin@u.arizona.edu.

Journal of Modern Applied Statistical Methods
May 2016, Vol. 15, No. 1, 2-11.

ISSN 1538 − 9472

Invited Article

The Composite Hypothesis Contrast Procedure: A Novel Sequential Multiple-Comparison Approach

Joel R. Levin
University of Arizona
Tuscon, AZ

Ronald C. Serlin
University of Wisconson-Madison
Madison, WI

The sequential composite hypothesis contrast multiple-comparison procedure is introduced for comparing two treatment conditions with one or two control conditions on one or two outcome measures. The procedure deserves consideration insofar as its power advantage over other commonly applied multiple-comparison methods can be sizable.

Keywords: Multiple-comparison procedures, sequential hypothesis testing, logical implications, comparison of means, analysis of variance contrasts

In the course of a recent research investigation—a single-case intervention study conducted by Hwang, Levin, and Johnson (2016)—we stumbled upon an interesting data-analysis situation that was reminiscent of one that had been considered a generation ago (Levin, Serlin, & Seaman, 1994). To summarize the take-home message of that 1994 article: Starting with a univariate $K = 3$ independent means one-way layout, we demonstrated that: (a) When an initial omnibus hypothesis test (of, for example, "All μ_k are equal") is rejected based on a Type I error probability of α, (b) if any sub-hypothesis subsumed by the rejected hypothesis is tested at α, then (c) the resulting familywise Type I error probability (α_{FW}) associated with entire set of tested hypotheses is equal to α.

The assertion follows, chronologically, from Fisher's (1935) least significant difference (LSD) protected multiple-comparison procedure when applied in a three-mean context; Fletcher's (personal communication, October 3, 1981) perceptive insights about that particular application of the procedure; Shaffer's (1986) introduction to, and cogent discussion of, the notion of logical implications

Dr. Levin is a Professor Emeritus in the Department of Educational Psychology. Email him at: jrlevin@u.arizona.edu. Dr. Serlin is a Professor Emeritus in the Department of Educational Psychology. Email him at rcserlin@wisc.edu.

of subsumed hypotheses; and the Monte Carlo simulation demonstrations of Seaman, Levin, and Serlin (1991), Zhou and Levin (2004), and others.

Consider a snapshot of logical implications in terms of controlling α_{FW} at α through Fisher's two-stage LSD procedure applied to a one-way ANOVA test of the equality of three independent means, μ_A, μ_B, and μ_C. In that situation, there are only three possible configurations of the three population means: (1) all differ from one another; (2) all are equal; and (3) two means are equal but they differ from the third mean. Let us consider each of these possibilities in turn, in the context of performing an omnibus one-way ANOVA *F*-test based on $\nu_1 = 2$ and $\nu_2 = N - 3$ degrees of freedom.

In Stage 1, the researcher conducts the omnibus *F*-test of H_0: All μ_k are equal. If, and only if, that hypothesis is rejected, the researcher proceeds to Stage 2 and applies a *t*-test to whichever mean differences (i.e., pairwise or complex contrasts) are of interest, each with a Type I error probability of α. If all population means differ, as in (1) above, and the omnibus-test hypothesis is rejected, then no Type I error can be made in the subsequent set of multiple comparisons because a Type I error can occur only when the means being compared are equal. Note that, in theory only, the researcher could declare that all means differ from each other without even conducting formal *t*-tests of the differences. Similarly, if the omnibus-test hypothesis is not rejected, no Type I error is made because the error incurred would be a Type II error.

If all population means are equal, as in (2) above, then the Stage 1 omnibus *F*-test provides the required α_{FW} control of the hypothesis tested. If the hypothesis is not rejected, that is a correct decision, no Type I error is committed, and no Stage 2 multiple comparisons are examined. If, on the other hand, the hypothesis is rejected, then a Type I error has been made with probability α. In that case, in Stage 2, any comparisons of interest can subsequently be examined because, with "familywise" referring to "at least one", the Type I error for the family has already been made and so it doesn't matter whether one, two, or a dozen more occur. Note that, in theory only, one could again declare that all means differ from one another without the formal *t*-tests.

Finally, if only two population means are equal, as in (3) above, if the Stage 1 omnibus hypothesis is not rejected then that again is a Type II error and the process is terminated. If, however, the hypothesis is rejected, then that is a correct decision and no Type I error has been made. Moreover, insofar as there is only one pair of means that are equal, there is only one opportunity for committing a Type I error in the subsequent set of Stage 2 *t*-tests. Thus, if each test is conducted based on a Type I error of α, then α_{FW} is also equal to α.

After detailing the underlying basis for the Fisher LSD procedure in the one-way ANOVA three-mean case, Levin et al. (1994) provided several extensions to other $v_1 = 2$ degree-of-freedom hypothesis-testing situations (e.g., main effects and interactions in 3×2 factorial designs, χ^2 tests in 3×2 contingency tables, Hotelling's T^2 two-group or MANOVA with two dependent variables). It is important to note that the same familywise Type I error control for the Fisher procedure does not hold for $K > 3$ or $v_1 > 2$ situations, even though Shaffer's (1986) logical implications and sequential testing procedures do (Levin et al., 1994; Seaman et al., 1991). Subsequently, similar sequential-testing logic associated with Scheffé's (1970) modified multiple-comparison procedure (Klockars & Hancock, 2000) was illustrated and extended by Zhou and Levin (2004) to hypothesis-testing situations with multiple independent or dependent variables (e.g., tests of P partial regression coefficients, K-group MANOVA with P dependent variables).

The Composite Hypothesis Contrast Procedure

In what follows, a novel sequential testing approach is proposed that is fundamentally different from both the Fisher LSD procedure and the planned Bonferroni-type procedures that were comprehensively reviewed by Shaffer (1986), Seaman et al. (1991), and Levin et al. (1994). Yet, the present approach obeys precisely the same type of successive logical implications that was just presented for the Fisher LSD procedure as applied to $v_1 = 2$ hypothesis-testing situations. With this new approach, a test of a single degree-of-freedom comparison (what we have termed a "composite hypothesis contrast") serves as a Stage 1 screening device, which, if proven to be statistically significant, leads directly to a set of logically implied α_{FW}-controlled additional contrasts. The procedure is so named because it essentially provides a framework for testing two linked hypotheses, first in combination and then individually.

The utility of the Stage 1 test of the composite hypothesis contrast is the same as that of initial omnibus tests associated with conventional multiple-comparison procedures, including those of Fisher (1935), Scheffé (1970), and Tukey (1953), among others. Specifically, if the Stage 1 test is statistically significant, it allows for α_{FW}-controlled follow-up testing of two focal hypotheses of interest. The fundamental assumption underlying application of the procedure is that two different experimental conditions are associated with similar differences or effects on the outcome measure(s), relative to other control or comparison conditions. Consider the approach for a few different comparison-of-means situations by beginning with a one-way layout with three independent conditions and a single

dependent variable, as would be applicable for the Fisher LSD procedure that we have been considering. Although the following discussion assumes equal sample-size situations, special comments on unequal sample sizes are included in the final part of the article.

Design 1: Three Conditions, One Outcome Measure.

In the three-condition case with two experimental conditions and one control condition, it is posited that the difference between each experimental condition and the control condition is of a comparable magnitude and in the same direction – but see the addendum that follows. (As an aside, the following discussion could alternatively assume that there is one experimental condition and two control conditions.) In the first stage of the procedure, the two experimental means are combined (i.e., averaged) and tested against the control mean as a composite hypothesis contrast based on a Type I error probability of α, via a *t*-test with the MS_W based on $\nu = N - K$ serving as an estimator of the population within-group variance. If statistically significant, in the second stage the two experimental conditions' means are separately compared with the mean of the control condition, each based on a Type I error probability of α. It is suggested both the composite hypothesis contrast and the follow-up separate contrasts typically be conducted as one-tailed tests insofar as a researcher would likely not be adopting this procedure without a solid rationale for and understanding of the direction of the treatment effects.

With α_{FW} controlled through logical implications, the procedure affords an efficient alternative to standard procedures for assessing both the aggregated and separate effects of the two experimental conditions. Specifically, the logical implications here are as follows: (1) If, in the population, either of the two experimental means differs from the control mean, then no Type I error is made with the Stage 1 test. Thus, if the Stage 1 hypothesis is rejected, then at most only one Type I error will be made with the two Stage 2 tests. (2) If, in the population, there is no difference between either of the two experimental means and the control mean, then a rejected Stage 1 hypothesis is a Type I error and, following the familywise Type I error concept, it does not matter whether zero, one, or two additional Type I errors occur during the Stage 2 testing.

Addendum. If (1) the outcome measure represents an interval scale, and (2) none of the to-be-described transformed data will fall beyond the measure's attainable upper or lower limits, then predicted experimental vs. control effects in opposite

directions can also be accommodated in the first stage test of the composite hypothesis contrast. For example, suppose it is predicted that the mean of one experimental condition will be higher than the control condition mean ($\mu_{E1} > \mu_C$) and the mean of the other experimental condition will be lower ($\mu_{E2} < \mu_C$). Further suppose that the actual sample means are in the predicted directions, with E1 exceeding C by 10 points and C exceeding E2 by 8 points. In that case, the E2 data could be transformed for the Stage 1 test by adding a constant of 16 (2×8) to all of the scores in that condition. As a result, the E2 mean will now be 8 points above the C mean, rather than 8 points below it, and the E1 and E2 means could be meaningfully combined for the composite hypothesis contrast test in the manner that was just described.

Design 2: Four Conditions, One Outcome Measure.

The composite hypothesis contrast procedure can be applied to test for differences involving four condition means in a manner similar to what was detailed for the three-condition case. Consider, for example, a study with two experimental conditions (E1 and E2) and two control conditions (C1 and C2). In addition, each experimental condition is conceptually linked to its own control condition: (e.g., E1 is linked to C1 and E2 is linked to C2). The researcher is testing for two similar treatment effects, one based on an ultimate comparison of E1 and C1 and the other based on an ultimate comparison of E2 and C2. The omnibus composite hypothesis contrast is initially tested at α in Stage 1 based on a comparison of the average of the E1 and E2 means with the average of the C1 and C2 means. If statistically significant, in Stage 2 the two separate contrasts are each tested at α, with the familywise Type I error rate controlled at α via logical implications analogous to those described for the three-group situation. In that regard, it is important to note that additional comparisons (e.g., of E1 and C2 or of E2 and C1) are not allowed as they would inflate the specified familywise Type I error rate.

Design 3: Two Conditions, Two Outcome Measures.

Now suppose that there are two conditions, experimental and control, and two different outcome measures of interest, X and Y. Moreover, it is assumed that similar treatment effects will be manifested on X and Y. Following the rationale of Marascuilo and Levin (1983) for creating an equally weighted linear combination of separate dependent variables by standardizing and adding (or averaging) them, a researcher could do the same here. In Stage 1, the composite hypothesis contrast procedure would initially compare the experimental and control condition on their

respective mean linear combinations (here, averages) of the X and Y measures, either standardized or unstandardized, depending on how comparable the two measures are assumed to be, based on a Type I error probability of α. If statistically significant, by logical implications in Stage 2 the experimental and control conditions means could be compared on the original X and Y outcome measures separately, each based on α, and thereby controlling α_{FW} at α.

Design 4: Four Conditions, Two Outcome Measures

A situation that incorporates aspects of Designs 2 and 3 was implemented in the previously cited Hwang et al. (2016) study where, in the context of a single-case crossover design (Levin, Ferron, & Gafurov, 2014), four different learning strategies (two experimental and two control) were predicted to have similar effects on two different outcome measures. Moreover, in that single-case design, the outcome measures of interest were the amounts of change/improvement between the baseline (A) phase and the intervention (B) phase of the study. In Stage 1 of the present statistical procedure, based on $\alpha = 0.05$, a one-tailed test of the composite hypothesis contrast proved to be statistically significant ($p = 0.020$). This result indicates that the composite hypothesis contrast (consisting of the two combined experimental strategies vs. the two combined control strategies), as applied to the change on the averaged two outcome measures, represented a detectable effect that was in the predicted direction. In Stage 2, for the two strategies' "comparison of change" tests on the two separate outcome measures, each at $\alpha = 0.05$, although both experimental strategies yielded effects that were in their expected directions, one of these was reasonably large and statistically significant ($p = 0.012$) while the other was considerably smaller and not statistically significant ($p = 0.087$).

The Dangers Lurking Beneath: Power Considerations

Just because the composite hypothesis contrast procedure can be implemented does not indicate that it is statistically advantageous or optimal to do so, relative to alternative α_{FW}-controlled multiple-comparison procedures that could be conducted instead. In particular, statistical power considerations would be advised when determining whether or when to use this approach.

Consider, for instance, the hypothetical examples presented in Table 1. There it is found that with a three-mean effect size defined as $f^2 = \omega^2/(1 - \omega^2)$, both where f is held constant at 0.471 in Parts A and B of Table 1 and as a general rule: (1) when the two averaged experimental means are equal and different from the control

mean (Panel A), Stage 1 of the present composite hypothesis contrast (CHC) approach overpowers at least three of its would-be competitors, namely, Fisher's

Table 1. Stage 1 powers for Fisher's LSD and the composite hypothesis contrast procedure, as well as powers to detect the larger of the two pairwise comparisons for the Holm-Bonferroni and Dunnett Procedures

A. Two means (E1 and E2) equal, each different from the third mean (C) by 1σ; three-mean effect size given by f = 0.471

n	Fisher	Holm (2T)	Holm (1T)	Dunnett (2T)	Dunnett(1T)	CHC (2T)	CHC (1T)
10	0.58	0.46	0.58	0.47	0.60	0.70	0.81
15	0.78	0.66	0.76	0.68	0.78	0.87	0.93
20	0.90	0.80	0.87	0.81	0.88	0.95	0.98

B. All means different in steps of 0.577σ (E1 > E2 > C); three-mean effect size given by f = 0.471

n	Fisher	Holm (2T)	Holm (1T)	Dunnett (2T)	Dunnett(1T)	CHC (2T)	CHC (1T)
10	0.58	0.59	0.70	0.60	0.72	0.58	0.70
15	0.78	0.80	0.87	0.80	0.88	0.76	0.85
20	0.90	0.91	0.95	0.90	0.95	0.87	0.93

Note: CHC = the present Composite Hypothesis Test; 2T = two-tailed test; 1T = one-tailed test

LSD Stage 1 omnibus test, along with Holm's (1979) sequential Bonferroni procedure and Dunnett's (1955) "each vs. one" multiple-comparison procedure applied to the larger of the two second-stage experimental vs. control comparisons; and (2) when the three means are more equally separated within the three-mean interval (Panel B), the one- and two-tailed test powers of the CHC approach are only slightly lower than those of the corresponding Holm and Dunnett powers, with the CHC approach's one-tailed powers still remaining higher than those of Fisher's LSD test.

In a previous study, Serlin and Mailloux (1999) investigated the analysis of designs with two conditions and two outcome measures, analogous to Design 3 above. They added together the two standardized outcome measures to form a composite that is similar to the composite that was described earlier here. Consistent with the our power results and conclusions, Serlin and Mailloux found that, if the univariate effect sizes associated the two measures are similar, with the smaller being at least half or more in size as the larger, then the Stage 1 screening test on the composite outcome measure followed by univariate tests in Stage 2 (as was presented here) is a more powerful procedure than both the multivariate Hotelling T^2 test and either Holm's (1979) or Shaffer's (1986) "sequentially rejective" procedures. Consequently, we have good reason to believe that, in the

present four-group application described earlier, that if the smaller of the separate E-C comparisons is at least half the size of the larger, the composite hypothesis contrast approach will also be more powerful than the alternative multiple-comparison procedures that were considered here.

Thus, there is a trade-off between the increased power resulting from the composite hypothesis contrast procedure based on the average of two equal or near-equal experimental means and reduced power resulting from a shrunken composite as the two averaged experimental means get further and further apart. In fact, we have determined that, as long as the ratio of the smaller to the larger experimental mean is at least 0.50, then as far as statistical power is concerned the CHC approach would likely be the hypothesis-testing method of choice in this three-mean situation. It is important to note nonetheless that the just-reported powers are not directly comparable. Those associated with the CHC and Fisher's LSD are Stage 1 omnibus test powers and those of Holm and Dunnett are Stage 2 powers for the larger of the two contrasts of interest. Yet it can be concluded that, because the Stage 2 critical values for the CHC are smaller than those for either Holm or Dunnett, if the CHC Stage 1 hypothesis is rejected, then the Stage 2 contrasts will be detected more often with the former procedure than they will with the two latter procedures.

Caveat

When constructing the composite hypothesis contrast, one must exercise caution in calculating the combined group mean in the case of unequal sample sizes, lest one fall prey to confounding due to what is known as Simpson's (1951) paradox. The paradox is perhaps easiest to envision as resulting from a third-variable influence in a two-way layout, wherein the unequal sample sizes are considered a function of a factor not considered in the design. In the earlier discussed Designs 2 and 3 with four conditions and one or two dependent variables, if weighted-by-sample-size means are used to form the Stage 1 composite hypothesis contrast, it is easy to show that, even if the E1 and C1 means were equal, as were the E2 and C2 means, then the combined E and C means in the composite could differ, in which case the logical implications required for the validity of the method do not hold. The solution, of course, would be to create the composite using unweighted means (i.e., the simple average of the E1 and E2 means minus the simple average of the C1 and C2 means).

Conclusion

The two-stage composite hypothesis contrast procedure is not a statistical panacea for all researchers in all multiple-comparison situations. It may, however, represent a useful statistical tool for some researchers in the situations for which it was intended, typically where two experimental treatments are expected to produce comparable effects (relative to one or two control conditions) on one or two outcome measures. The procedure is recommended for those situations because it provides a straightforward, more powerful statistical alternative to other commonly applied multiple-comparison methods.

References

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association, 50*(272), 1096-1121. doi: 10.1080/01621459.1955.10501294

Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65-70. Available from http://www.jstor.org/stable/4615733

Hwang, Y., Levin, J. R., & Johnson, E. W. (2016). Pictorial mnemonic-strategy interventions for children with special needs: Illustration of a multiply randomized single-case crossover design. *Developmental Neurorehabilitation*. Advance online publication. doi: 10.3109/17518423.2015.1100689

Klockars, A. J., & Hancock, G. R. (2000). Scheffé's more powerful *F*-protected post hoc procedure. *Journal of Educational and Behavioral Statistics, 25*(1), 13-19. doi: 10.3102/10769986025001013

Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2014). Improved randomization tests for a class of single-case intervention designs. *Journal of Modern Applied Statistical Methods, 13*(2), 2-52. Retrieved from http://digitalcommons.wayne.edu/jmasm/vol13/iss2/2

Levin, J. R., Serlin, R. C., & Seaman, M. A. (1994). A controlled, powerful multiple-comparison strategy for several situations. *Psychological Bulletin, 115*(1), 153-159. doi: 10.1037/0033-2909.115.1.153

Marascuilo, L. A., & Levin, J. R. (1983). *Multivariate statistics in the social sciences: A researcher's guide*. Monterey, CA: Brooks/Cole.

Scheffé, H. (1970). Multiple testing versus multiple estimation: Improper confidence sets, estimation of directions and ratios. *Annals of Mathematical Statistics, 41*(1), 1-29. Available from http://www.jstor.org/stable/2239715

Seaman, M. A., Levin, J. R., & Serlin, R. C. (1991). New developments in pairwise multiple comparisons: Some powerful and practicable procedures. *Psychological Bulletin, 110*(3), 577-586. doi: 10.1037/0033-2909.110.3.577

Serlin, R. C., & Mailloux, M. (1999, April). *An empirical comparison of three methods for performing univariate analyses with multivariate data*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Québec.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association, 81*(395), 826-831. doi: 10.1080/01621459.1986.10478341

Simpson, E. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological), 13*(2), 238-241. Available from http://www.jstor.org/stable/2984065

Tukey, J. W. (1953). *The problem of multiple comparisons* (Unpublished manuscript). Princeton University, Princeton, NJ.

Zhou, X., & Levin, J. R. (2004). A note on extending Scheffé's modified multiple comparison procedure to other analysis situations. *Journal of Modern Applied Statistical Methods, 3*(2), 432-442. Retrieved from http://digitalcommons.wayne.edu/jmasm/vol3/iss2/15/