Journal of Methods and Measurement in the Social Sciences
Vol. 2, No. 2, 63-79, 2011

Simple, Powerful Statistics: An Instantiation of a Better 'Mousetrap'

Mark Roberts
British Columbia, Canada

R.S. Rodger fully developed, more than three decades ago, probably the most powerful methodology which exists for detecting real differences among population means (μ's) following an analysis of variance. Since it is a *post hoc* method, a theoretically infinite number of potential statistical decisions may be considered, but Rodger's method limits the final number of decisions to a single set which contains exactly J-1 (i.e., v_1, the number of means in a study minus one) of them. It also constrains the number of these J-1 decisions that may be declared statistically "significant." Rodger's method utilizes a decision-based error rate, and ensures that the expected rate of rejecting null contrasts that should not have been rejected (i.e., the type 1 error rate) will be less than or equal to either five or one percent, regardless of the number of contrasts examined by a researcher prior to finally deciding upon the scientifically optimal set of decisions.

The greatest virtue of Rodger's method, though, is not its considerable power, but its explicit specification of the magnitude of the differences that the researcher will claim to exist among the population parameters. The implied true means that this method calculates are the theoretical population μ's that are logically implied, and mathematically entailed, by the J-1 statistical decisions that the researcher has made. These implied true means can assist other researchers in confirming or disconfirming population parameter claims made by those who use Rodger's method. A free computer program (SPS) that instantiates Rodger's method, and thereby makes its use accessible to every researcher who has access to a Windows-based computer, is available from the author.

Keywords: Rodger's method of *post hoc* analysis, decision-based error rate expectation Eα, non-traditional test criterion *F*[Eα], implied true means (μ's), SPS computer program

Ralph Waldo Emerson's saying, "build a better mousetrap and the world will beat a path to your door," seems not as commonly heard now. This is perhaps so because the saying may not be true for even a majority of Emersonian mousetrap equivalents. R.S. Rodger invented a significantly better procedure for statistical decision-making following an analysis of variance (ANOVA) and introduced it in seven articles published in the prestigious-enough British Journal of Mathematical and Statistical Psychology in the 1960's and 70's. Since then, it has languished and is not widely known or frequently used. [1]

[1] An opinion to the contrary was expressed by Williams, Frame, & LoLordo in their 1992 article: "We chose Rodger's method because it is the most powerful post hoc method available for detecting true differences among groups. This was an especially important consideration in the present experiments in which interesting conclusions could rest on null results. Although *Rodger's method is commonly used …*" (p. 43, emphasis added).

Rodger's Method: Implied True Parameters from Decisions Made *Post Hoc*

What accounts for the superiority of Rodger's method? The short answer is that it uses a decision-based (per contrast) type 1 error rate and new critical *F* values, which jointly produce this effect: "Rodger's approach ensures that statistical power does not decline (and even increases) with increasing numerator degrees of freedom" (Delamater, Campese, & Westbrook, 2009; p. 228). This stands in direct opposition to the characteristics of traditional experiment-wise error rates and commonly used *F* values or studentized range values for other *post hoc* procedures (e.g., those of Scheffé, Tukey, and Newman-Keuls). Using traditional *F* table values when the null hypothesis is genuinely false, produces a serious *decrease* in the probability (β) of detecting a greater-than-zero treatment effect of fixed size with increasing numerator degrees of freedom.

This problem with traditional experiment-wise error rates is likely most easily recognized in the context of a factorial analysis, where there is greater power to detect a non-zero treatment effect in the I main effect than the J main effect (assuming there are fewer I than J treatment levels), which itself has greater power to detect a fixed-sized effect than the I x J interaction. In the most important article about his method (by my reckoning), Rodger (1974) says: "this feature is supposed to be well known, since it is a direct result of the unequal number of replications in the different effects. ... When it is said that factorial analyses will inform us whether or not interactions exist, the tongue should be held in the cheek!" (p. 194).

By adopting a decision-based type 1 error rate and calculating new decision-wise *F* values to be used when making *post hoc* statistical decisions, Rodger's method avoids the problems associated with traditional experiment-wise *F* values. To accomplish this, Rodger begins (as did Scheffé) by noting that the analysis of variance is a procedure that partitions the overall, between-groups variance into J-1 (i.e., v_1, the number of means being analyzed minus one) completely independent components. Consequently, when J-1 mutually orthogonal contrasts are constructed following an ANOVA, the sum of the *F* values for each of the individual contrasts in this set will necessarily be equal to the overall ANOVA variance ratio (i.e., the omnibus *F* value, denoted here as F_m).

Unfortunately, when partitioning the overall ANOVA between-groups variance into independent components (and J > 2), it is theoretically possible to do this in an infinite number of ways (i.e., that many, infinitesimally different from one another, sets of J-1 mutually orthogonal contrasts could be constructed). Scheffé's decision rule permits an infinite number of contrasts to be declared significantly different from zero whenever F_m for a particular study is equal to or greater than the traditional critical F value. Rodger chose a different approach, and instead considered how an appropriate limit (r) could be placed on the number of declarations of statistical significance. The fact that the expectation of r/v_1 is a linearly decreasing function of the numerator degrees of freedom (v_1) when the usual F table values are used (alluded to in the second paragraph of this article), led Rodger to construct new F values that keep the expectation of r/v_1 at a constant .05 or .01.

These new critical F values (denoted $F[E\alpha]; v_1, v_2$ to distinguish them from the traditional values of $F\alpha; v_1, v_2$) appear in separate tables for $E\alpha$ = .05 and $E\alpha$ = .01 in Rodger (1975a). Rodger's method uses the tabled $F[E\alpha]$ value for a given study in two ways. It is first used to set a limit on the maximum number of rejected null contrasts (i.e., declarations of statistical significance) that Rodger's method permits. This limit is

$$r = [F_m / F[E\alpha]; v_1, v_2] \leq v_1;$$

or, in plain English, the maximum number of rejectable contrasts equals the integer value of the ANOVA variance ratio divided by Rodger's critical F value, but cannot be permitted to exceed J-1. The second use for the critical $F[E\alpha]$ value for a particular study is to determine which specific contrasts are rejectable. When $F_m \geq F[E\alpha]; v_1, v_2$, Rodger's method permits as many as r contrasts to be rejected (though that many may not always be scientifically interpretable), and any contrast with an F value equal to or greater than $F[E\alpha]; v_1, v_2$ is a rejectable one.

Except when $v_1 = 1$ (i.e., there are only two means), Rodger's $F[E\alpha]$ values are lower than the traditional $F\alpha$ values. Using $F[E\alpha]$ values in the two ways mentioned in the previous paragraph ensures that when r contrasts are actually rejected in the set of decisions that the researcher ultimately adopts, the expected rate of null-contrast rejections will be $E\alpha$ (i.e., .05 or .01) when all null contrasts are true. A number of researchers have used the greater power to detect non-zero treatment effects that Rodger's $F[E\alpha]$ tables afford, and this can be done without further analysis. Utilizing Rodgerian *post hoc* decision-making in this rather limited way is certainly an improvement over the procedures that researchers typically use, and requires nothing more than looking up the $F[E\alpha]$ values and using them instead of the traditional $F\alpha$ values to find r or fewer null-contrast rejections. [2]

Using Rodger's full method, though, requires considerable spreadsheet expertise or a computer program that correctly implements it. Until now,

[2] Peter Urcuioli, for example, has used this aspect of Rodger's methodology in about a dozen published articles in the past decade (e.g., 2008), and in many others going back to 1981.

only Rodger's own program (IMPLY, written in FORTRAN for use on mainframe computers) has done this, and it was not widely disseminated. My SPS (Simple, Powerful Statistics) computer program is written in Visual Basic for use on a Windows-based personal computer. The SPS implementation of Rodger's method requires that its users make J-1 statistical decisions that are constrained in the following ways.

First, these statistical decisions are expressed by assigning numbers (contrast coefficients) to the means that indicate how they are to be compared with one another. Each of the J-1 decisions is required to be stated in the form of contrasts, which simply means that the sum of the coefficients for each contrast (entered across one row in the contrast matrix) must equal zero. The simplest contrasts are comparisons (with coefficients 1 and -1) that assess the magnitude of the difference between two selected means. Means that are not part of a statistical decision are given contrast coefficients of zero, and contrasts of any degree of complexity are permitted. The full contrast matrix has J-1 (the number of contrasts) rows and J (the number of means) columns, and each cell contains either a zero or a non-zero number.

Second, when the *F* value of a contrast equals or exceeds Rodger's critical *F*[Eα] value and that contrast is included in a "decision set," it counts as one of the r rejectable contrasts that are permitted. [A decision set is comprised of a matrix of contrast coefficients for J-1 contrasts, along with a vector of δ_h values (discussed below) that have been decided for those contrasts, and a vector of μ symbols.] If r is less than J-1, as it frequently will be, then at least υ_1 - r of the decisions will necessarily be accepted null contrasts. Scientists often have good reason to declare that no theoretically meaningful differences exist among two or more populations represented by two or more sample means, but more is needed to justify such declarations than the mere statistical non-significance of the *F* values of particular contrasts.

What warrants declarations that no meaningful differences exist among two or more population means is this: 1) the researcher specifying a theoretically meaningful effect size (denoted by Rodger as a g value), and 2) using this information to ensure that a sufficient number of subjects are included in the study to have enough power to detect non-zero effects of that size. SPS makes it a simple matter to use the tables that Rodger (1975b; 1978) provided which permit the determination of the sample size needed to set the null-contrast rejection rate at a desired level of Eβ (say, .95) when using one-stage (or, possibly two-stage) sampling. In other words, this determination of the sample size required to detect theoretically specified non-zero effect sizes needs to occur at the design stage, so that the accepted null contrasts in the decision sets that are considered during the analysis stage are defensible.

Third, all of the contrasts in a decision set must be linearly independent of one another, and will preferably be mutually orthogonal; i.e., be characterized by the strong form of linear independence that occurs when the cross-products of the contrast coefficients for each of the (J-1) (J-2) / 2 pairs of contrasts in the decision set sum to zero.

Rodger's method is predicated on evaluating such appropriately constrained J-1 statistical decisions as a single set. There is a unique aspect of this method that Rodger summarizes in a single sentence in the abstract of his 1974 article: "Theoretical means are deducible from decisions for υ_1 mutually orthogonal contrasts and, if rejected null contrasts are given suitable non-zero values, there is no ambiguity about the theoretical means" (p. 179). These non-zero values for the rejected null contrasts (g values) are the "scale-free" component of the linear, non-central parameter δ_h (formula 9 in Rodger, 1975b), and they are specified by the researcher to indicate the size of the non-zero effects the experiment or study was designed to detect. [3] Accepted null contrasts are assigned a g value of zero (which signifies a negligible treatment effect), and the δ_h value that is decided for each contrast is the product of the g value for a contrast and the sum of the squared contrast coefficients for that contrast. As if by magic, at least to those of us who are mathematically challenged, Rodger's method takes the contrast matrix and δ_h values and works out what population parameters (μ's in this case) are mathematically necessitated by the specific decisions that the researcher has made. Unless two-stage sampling has been used (which permits purely numeric decided values), these implied true means have to be expressed in units of the unknown population standard deviation (σ). Nevertheless, the implied true means are a very precise expression of the outcome of the study – these theoretical population parameters quantify the magnitude of the differences that the researcher is claiming to exist among the population means (μ's).

This does not deny the value of directly reporting the sample observed effect sizes, as is now commonly advocated – it simply quantifies what the investigator believes the "true effects" are and allows more precise checks on those claims by others. By reporting the g and Eβ values used in the design of a study, and the implied true means that constitute the study's outcome, valuable information is provided that can guide future researchers on that topic in deciding upon the sizes of effects that their experiments will be designed to evaluate.

[3] "Obviously we do not know the true value of g_h in any statistical investigation; we can state only what value we hope to detect with probability β. It is sometimes possible to specify a value for g_h which seems reasonable in the light of previous research and theory, but in the absence of such information it is not unreasonable to set g_h ±1, which is a moderately large effect that may be detected with good power by small samples" (Rodger, 1974; p. 189). SPS uses this suggested value of 1 as its default for the g values.

Whenever the ANOVA variance ratio (F_m) for a particular study is equal to or greater than Rodger's critical F[Eα] value, it is always mathematically possible to construct mutually orthogonal sets of contrasts with r rejected and v_1 - r accepted null contrasts. It is also possible, though, that the contrast coefficients for the decisions in these sets may be quite complicated ones that are not readily interpretable scientifically. Rodger has consistently maintained that scientific meaningfulness is paramount and always trumps mathematical possibility when it comes to constructing mutually orthogonal decision sets with r rejected null contrasts. If faced with the absence of a scientifically-sensible, orthogonal decision set that contains r rejected contrasts, the researcher has two available options. The first is to instead adopt a scientifically meaningful decision set that has r rejected contrasts that are linearly independent but not mutually orthogonal. The second option is to adopt a scientifically meaningful decision set that has fewer than r rejected contrasts. The implied true means are always computable by the user of Rodger's method provided that the contrasts in the decision set are linearly independent of one another and a non-zero g value is specified for each (and at least one) rejected null contrast.

A Simple Illustration of Rodger's Method

In this section I will basically reiterate, and try to make more clear, some of the important aspects of the foregoing discussion of Rodger's method by using a numerical example. A few new points will also be made. Although it may not be good statistical practice to pick up someone else's data and re-analyze it, I will frame my example in this way so that it can be devoid of any subject-matter meaning and the focus can remain solely on the numbers.

Assume that we happen to find means and standard deviations from a discarded experiment that had N = 8 subjects in each of three independent groups: 1) 4.688 & 0.567, 2) 4.825 & 0.889, and 3) 5.475 & 0.486. When this information is processed by a suitable computer program (e.g., SPS), a one-way analysis of variance summary table may be viewed. In this case, the overall F value (F_m) is 3.147 and the traditionally used critical $F_{.05}$ value with 2 and 21 degrees of freedom is 3.47. Obviously, F_m is not large enough to be judged statistically "significant" and the unknown researcher's conclusion must have been that nothing of interest was found (hence, since statistical decisions are always made about population parameters rather than obtained sample values, $\mu_1 = \mu_2 = \mu_3$ is plausible).

However, if a more powerful procedure were used, namely Rodger's, the researcher would have come to a very different conclusion about his/her experiment. When Rodger's method is used and r is greater than zero, as it is here, it is always possible to find r rejectable null contrasts across the population parameters (μ's in this case). Rodger's critical *F*[.05]; 2, 21 value is 2.739, so Rodger's method can reject r = [3.147 / 2.739] = 1 null contrast in a decision set with two contrasts, and υ_1 - r (i.e., 2 - 1 = 1) of the contrasts in the decision sets constructed for this illustration will be a single, accepted null contrast. In any study, the difference between one rejectable null contrast and none is surely "a difference which makes a difference" (this phrase was Gregory Bateson's definition of information).

There is a colloquial (and not literally true) sense in which Rodger's method also ensures that each one of the as many as r rejected null contrasts the researcher includes in the adopted decision set (despite perhaps contemplating the scientific merit of numerous contrasts in many decision sets) will have at most an .05 or .01 chance of having been rejected in error. In the same sense, the probability of it being a type 1 error will be less than five or one percent for each rejected null contrast if the adopted decision set contains fewer than r rejections. The correct way of putting this was stated above – when Rodger's *F*[Eα] values are used and r null-contrast rejections are included in a set of J-1 decisions, "the expected rate of null-contrast rejections will be Eα (i.e., .05 or .01) when all null contrasts are true." When using Rodger's method over the long-run, the rate of rejecting null contrasts that should not have been rejected will be Eα. If, however, the one null-contrast rejection that is included in the decision set that will soon be adopted in this illustration is actually a true null, the probability of my committing a type 1 error when I reject it will be 100 percent. It is the long-run average of making this type of error that Rodger's method ensures will be no larger than five (or one) percent. It is important to remember than an error occurs, in the statistical context, if and only if a decision is made that a specified relationship among population parameters either is, or is not, equal to some number (usually, zero), and the opposite is true. Rodger's very sensible, and cogently argued, position is that statistical error rate should be based exclusively on those things in which errors may occur, and that (necessarily, by definition) can only be the statistical decisions that researchers make.

When the population parameters in any experiment are not all equal to each other, as is presumably true in this example, the question that naturally arises is: How, specifically, do the population means (or proportions, [4] or ranks) differ from one another? Part of the researcher's job is to make scientifically-informed statistical decisions and then make a claim about the answer to this question. A maximum of J-1 statistical

[4] See Rodger (1969) for the application of his method with independent proportions. In addition to providing a Rodgerian analysis of means, SPS also does this for proportions and for ranks. All three of these types of data may come from either independent or correlated groups.

decisions may be made without introducing contradiction or unnecessary repetition into a decision set. Repetition is sometimes a good thing, as I hope you can appreciate in this section of my article. But it is not desirable in a set of statistical decisions. In this context, repetition of previous information, and the far worse offense of contradicting oneself, may not even be recognized because doing so may require a greater ability than any of us possess to "see" logical implications that may have to be deduced from many concurrent decisions. Rodger's method prevents repetition, and contradiction, from entering into the statistical decision-making process by restricting the decisions that may be made to J-1 of them that are all at least linearly independent of one another. Consequently, just two statistical decisions that satisfy the linearly independent criterion will appear in each of the three decision sets that will be constructed and considered for possible adoption. To reiterate, no more than r = 1 decision of the J-1 = 2 decisions to be made will be a null-contrast rejection, so that we can know that the probability of a type 1 error will not exceed the selected Eα = .05 level.

As stated earlier, the simplest contrasts are comparisons of one mean with another. With the three sample means arranged in ascending order and denoted m_1, m_2, and m_3, the sum of the contrast coefficients (0, -1, and 1) times the sample means ($\Sigma c_j m_j$) for all three possible ways of comparing two means at a time, and the F values of these contrasts are: 1) $m_2 - m_1 = 0.137$; $F = 0.084$, 2) $m_3 - m_2 = 0.65$; $F = 1.881$, and 3) $m_3 - m_1 = 0.787$; $F = 2.757$. Only the third contrast comparing the largest mean with the smallest one has an F value that exceeds Rodger's critical F[.05] value of 2.739, so we can include that rejected null contrast in a decision set and thereby declare that $\mu_3 - \mu_1 > 0$ (i.e., $\mu_3 > \mu_1$). If someone were inclined to do something that Rodger's method precludes, and assert the truth of all three of these statistical decisions (since each one does have some statistical justification), contradiction will ensue. This follows from the fact that if $\mu_3 - \mu_2 = 0$ (contrast two) and $\mu_2 - \mu_1 = 0$ (contrast one) are accepted as being true, then that individual would be claiming both that $\mu_3 = \mu_2$ and that $\mu_2 = \mu_1$. These two decisions taken together logically require that $\mu_3 = \mu_2 = \mu_1$ and, since this is so, one cannot simultaneously, without contradiction, also accept the third contrast which states that $\mu_3 > \mu_1$. The dictates of logic need to be accorded priority over statistical possibility, especially when a statistical methodology permits such folly. Logic, or folly, will be embedded in the researcher's choice of what to claim.

Two obvious decision sets with J-1 = 2 decisions can be constructed from the three contrasts above as follows: 1) use the rejectable third contrast with the non-rejectable first contrast, and 2) use the rejectable third contrast with the non-rejectable second one. The two contrasts are not orthogonal to one another in either of these two decision sets, but they are linearly independent of one another so either set could be adopted. As already noted, the rejected null contrast that is included in both decision sets asserts that $\mu_3 > \mu_1$, the accepted (i.e., non-rejected) null contrast in decision set #1 asserts that $\mu_2 = \mu_1$, and the accepted null contrast in decision set #2 asserts that $\mu_3 = \mu_2$. The logically implied ordering of the three μ's in these two decision sets is thus as follows: set #1 says that $\mu_1 = \mu_2 < \mu_3$, and set #2 says that $\mu_1 < \mu_2 = \mu_3$. It is clear, however, that μ_2 is much closer to μ_1 than it is to μ_3 (since $m_2 - m_1 = 0.137$ while $m_3 - m_2 = 0.65$), so the pair $\mu_2 - \mu_1 = 0$ and $\mu_3 - \mu_1 > 0$ (i.e., decision set #1) will fit the data better.

Of course, there are lots of other possible decision sets for three means, including many that are mutually orthogonal. If we start with the rejectable contrast above (comparison three), we can construct a contrast that is orthogonal to it by taking the average of the lowest and highest means and compare that with the middle one; in other words, theoretically, $(\mu_1 + \mu_3) / 2 - \mu_2$. Contrasts, as previously noted, must have coefficients that sum to zero, and for this particular one they will be 0.5, -1, 0.5 or, if each is multiplied by 2 in order to make integer rather than decimal coefficients, 1, -2, and 1. For this contrast, $(m_1 + m_3)/2 - m_2 = 0.513$ and its F value is 0.390. This contrast (with coefficients 1, -2, 1) is orthogonal to the rejectable contrast (with coefficients -1, 0, 1), because the cross-products of the two sets of coefficients (1 × -1, -2 × 0, and 1 × 1) sum to zero. In addition, when all J-1 contrasts are mutually orthogonal (i.e., every contrast is orthogonal to every other one), the sum of the contrast F values will be the overall F value ($F_m = 2.757 + 0.390 = 3.147$). This is mathematically necessitated by the fact that, as stated earlier, "the analysis of variance is a procedure that partitions the overall, between-groups variance into J-1 ... completely independent components." The non-rejectable contrast here clearly states that μ_2 is somewhere between the other two, so this orthogonal, third decision set logically implies that $\mu_1 < \mu_2 < \mu_3$.

In addition to these three decision sets, many of the infinite number of not-so-simple sets could (but mercifully, won't) also be constructed. There are, after all, only two basic logical orderings of three population means that are not deemed to all be equal to one another: 1) two of the μ's are equal to each other and greater than or less than the third, or 2) all three μ's are different from one another and they are ordered in one of six possible arrangements. The number of possible patterns of arrangements is greatly expanded, of course, as the number of means in a study increases. But even in this simple example, the three decision sets that have been constructed imply three quite different orderings of the true population means for the data being considered.

How should we decide which of these three specific orderings is the most scientifically sensible one? In this instance we can't say anything about this because, despite having now found the raw scores that the stated descriptive statistics were calculated from (a fortuitous occurrence that facilitates analyzing the data with other computer programs and alternative procedures), we still know nothing about this experiment. But Rodger's fundamental answer to this question is that the researcher's understanding of theory and prior research in the domain that the experiment or study was designed to investigate should play a determinative role in making this decision. Ideally, I think he would say, anyone using his method should be able to construct multiple (preferably, mutually orthogonal) sets of contrasts that reflect theoretically interesting differences that might exist among the population parameters, and then exercise good judgment when deciding which set to adopt. In this example, the numbers that are being considered will remain decontextualized. Even so, Rodger's method provides important information that must be considered when choosing among competing sets of decisions.

With three different decision sets that imply quite different orderings of the population μ's, it is time to do the Rodgerian matrix magic (Rodger provides an illustration of how this is done for three means at equation 23 in his 1975b article, but this is impressive whether you understand its mathematical basis or not). Using a constant value of g = 1 (see footnote 3), the three sets of implied true means for the decision sets created above are: 1) -.47σ, -.47σ, .94σ; 2) -.94σ, .47σ, .47σ; and 3) -.71σ, 0σ, .71σ. As expected, these three sets of implied population means are very different. The statistical help that Rodger's method offers in the decision-set selection process consists of two measures of how closely associated the sample means from a study are with the implied true means for each decision set that will be considered for possible adoption.

One obvious measure of the "fit" between the sample means and implied true means (or ranks, or proportions) is provided by the Pearson correlation coefficient, which for the three decision sets in this example are: .986, .637, and .937. Another fit statistic is discussed in Rodger (1978, p. 169-170), which quantifies the amount of variation in the sample means that is not accounted for by the implied means. The closer this residual *F* value is to zero, the better the fit. The obtained residual *F* values for the three decision sets under consideration are equal to the non-rejected contrasts' *F* values: .084, 1.881, and .390. A relatively low fit residual value (or correspondingly, a high correlation between the sample and implied means) should be regarded as a necessary condition for concluding that a particular decision set that is being considered for possible adoption is optimal. Ordinarily, neither of these two assessments of fit can ever be taken as a sufficient condition for determining which is the best of the alternative decision sets vying for adoption consideration. Only in a highly improbable scenario such as this one should a researcher have nothing better than fit statistics upon which to base the choice regarding the set of decisions that will be asserted and interpreted.

From this simple illustration of Rodger's method of *post hoc* analysis of means, it should be apparent that statistical decisions have logical implications about the population parameters that are of interest in every experiment. In the simple case of three means illustrated here, it is easy to keep track of the logical implications of the contrasts included in each of the three decision sets. With as few as four means it starts to become rather difficult to "see" even the ordinal positioning of all four implied means prior to doing a Rodgerian analysis of those means. And with not many more than four means, it is difficult, if not impossible, to either intuitively or laboriously decipher how the contrasts in a decision set collectively affect the population parameters (e.g., the implied true means) that they logically imply. Usually, the only way to fully know what is logically implied by the statistical decisions that researchers make is to do the mathematical (i.e., matrix) operations that underlie Rodger's insight into how this can be achieved.

This illustration of Rodger's method is nearly finished, and much of what has thus far been discussed is summarized in Table 1. As noted above, the data analyzed there are for sample means m_j = 4.688, 4.825, 5.475 and standard deviations = .567, .889, .486, each based on N = 8 subjects, and yielding ANOVA F_m = 3.147. Rodger's (1975a) critical F[0.05]; 2,21 = 2.739, which makes the number of rejectable null contrasts r = [3.147/2.739] = [1.15] = 1.

Table 1
Three Possible Sets of Contrast Pairs and Their Implied μ_j

Set & contrast#	Contrast coeff. c_j	$\Sigma c_j m_j$	Contrast F value	Ordering of the μ's	Fit statistics corr.	resid.	Implied true μ_j
I (1)	-1, 0, 1	.787	2.757	$\mu_1 = \mu_2 < \mu_3$	.99	0.084	-.47σ, -.47σ, .94σ
(2)	-1, 1, 0	.137	0.084				
II (1)	-1, 0, 1	.787	2.757	$\mu_1 < \mu_2 = \mu_3$	.63	1.881	-.94σ, .47σ, .47σ
(2)	0, -1, 1	.650	1.881				
III (1)	-1, 0, 1	.787	2.757	$\mu_1 < \mu_2 < \mu_3$	.94	0.390	-.71σ, 0σ, .71σ
(2)	1, -2, 1	.513	0.390				

Note: The implied true means (μ_j) must be expressed in units of the unknown population standard deviation when the usual single stage of sampling is employed.

On statistical grounds alone, it is apparent that the contrasts in decision set II imply population means that are not well-fitted to the sample means obtained in this experiment. The second decision set should not be seriously considered for adoption. Either of the other two decision sets might be the best choice depending on what the researcher knows about the field that this experiment was investigating. In this illustration, that knowledge has been presumed to be non-existent, so decision set I seems the best choice.

What are the implications of this example for my assertion (in the first sentence of the article abstract) that Rodger's method is probably the most powerful *post hoc* procedure in existence? The single null-contrast rejection that was included in all three of the decision sets that were constructed for this example is the same one: $\mu_3 - \mu_1 = 0$. The data here originally belonged to someone else, and we can ask what other *post hoc* procedures that person could have employed that have sufficient power to similarly reject $\mu_3 - \mu_1 = 0$. When a one-way analysis of variance is performed on the raw scores for these three groups by a frequently-used statistical analysis program (SPSS), and all the tests on offer under the "Equal Variances Assumed" heading are selected, only Fisher's LSD (Least Significant Difference) test receives the honorific asterisk signifying that "the mean difference is significant at the 0.05 level." The LSD test does a "protected *t*-test" on as many comparisons as can be found, but only if $F_m \geq F.05, v_1, v_2$ (i.e., the traditionally used critical *F* value). When this criterion is met (and it isn't in this example, since $3.147 < 3.47$), Fisher's LSD test typically permits the researcher to declare a lot of null contrasts to be statistically “significant."

Consequently, it appears that Rodger's method is the only *post hoc* procedure that can credibly reject $\mu_3 - \mu_1 = 0$ for the data in this example. As previously stated, it does this by: 1) requiring that the researcher make exactly the same number of statistical decisions (i.e., v_1 = J-1) as the number of orthogonal components that an analysis of variance decomposes the overall variation among the sample means into, 2) allowing no more than r of these decisions to be null-contrast rejections (i.e., declarations of statistical significance), and 3) using Rodger's $F[E\alpha]; v_1, v_2$ values. The decision-based error rate that Rodger's method uses is responsible for the increased power that his method possesses, and it also ensures that the long-run average number of null-contrast rejection errors (type 1 errors) when using this method will be less than or equal to $E\alpha$ = .05 or .01. As an anonymous reviewer of an earlier version of this article put it: “lack of power in conventional post-hoc procedures ... is [largely attributable to] unneeded control for the family-wise [actually, experiment-wise] error rate." With Rodger’s method, a researcher is permitted unlimited access to *post hoc* data snooping, and, because a decision-based error rate is utilized (as is true of planned *t*-tests), also generally has more power to detect real differences among population parameters than can be obtained with other statistical procedures. [5]

I leave it to any interested statisticians to do the work necessary to conclusively establish whether or not Rodger's method is, as seems very likely, the most powerful *post hoc* procedure in existence. The most important part of the story being illustrated here, though, will be true regardless of the future consensus among statisticians about the "most powerful" title. If there were another *post hoc* procedure that was powerful enough to reject $\mu_3 - \mu_1 = 0$ in this example, the user of that (presumably non-existent) method could also claim that $\mu_1 < \mu_3$. But where does μ_2 fit into the picture? The answer is obvious – its placement relative to the other two μ's is entirely dependent on what the researcher chooses to claim. The statistical decisions that researchers are free to make (such as which accepted null contrast to use to complement the obvious, rejectable null contrast in this simple illustration) result in different orderings of the implied population μ's. It is not enough to merely say that some of the population means are unequal to one another. The researcher should specifically indicate in what way she/he believes they are unequal. That is exactly what *post hoc* testing is supposed to do – assist the researcher in making decisions regarding what to claim about the population parameters. There can be no doubt at all that Rodger's method is the only *post hoc* procedure in existence that explicitly provides the "true" population parameters that are logically implied, and mathematically entailed, by the statistical decisions that researchers make. Rodger's method will do this for up to 61 means, or ranks, or proportions as readily (though not quite as quickly, since more mathematical calculations are necessary) as it did so for the 3 means in this illustration.

The SPS (Simple, Powerful Statistics) Implementation of Rodger's Method

The raison d'être of the SPS computer program is to make Rodger's method of *post hoc* decision-making accessible to researchers. The

[5] Implicit in this sentence is the compelling reason for doing *post hoc* rather than planned tests in the first place. With *post hoc* decision-making it is possible to: 1) say something important about population parameters without having to know what you are going to say before collecting your data, and 2) preclude ever being in the unenviable position of merely reporting that what you intended to say isn't warranted. Quite apart from this, the power inherent in Rodger's method permits differences between population parameters to be found very economically. In his re-analysis of an experiment in which there were 10 subjects per group, Rodger claims that the difference of .37σ between two specific implied μ's should be regarded as a real difference in the population μ's (in that particular instance). He then notes: "To detect a difference $\mu_{12} - \mu_{22} = -0.37\sigma$ in a planned, two-sided *t* test with $\alpha = 1 - \beta = 0.05$ would require N = 190 in each sample" (Rodger, 1974; p. 197).

purpose of this article is to hopefully interest some people in using (or at least giving some further consideration to) Rodger's method. It thus seems reasonable to say a bit more about how easily Rodgerian statistical analyses can be performed with the SPS program.

The input data that SPS typically uses are raw scores from a study that have been saved in a comma-delimited or tab-delimited text file. When the raw scores in a datafile are read in, a one-way independent groups, or repeated-measures, or split-plot (mixed model) ANOVA is calculated in order to obtain the overall variance ratio (F_{m}) which, along with the retrieved critical F[Eα] value, is needed to calculate the maximum number of rejectable null contrasts (r). [6] Whether the data come from raw scores, or means and standard deviations from independent groups (as noted in the previous section of this article), the ANOVA summary table can, optionally, be displayed and printed.

The only somewhat difficult part of using Rodger's method is constructing sets of J-1 mutually orthogonal (or, minimally, linearly independent) contrasts and selecting the one that reflects the scientifically meaningful decisions about population parameters that the researcher wants to make. This is not surprising because, from the SPS program user's perspective, that's just about all there is to using Rodger's method. As just described, the dual aspects of this task have until now been fairly closely connected. Specifically, the researcher would previously: 1) consider alternative interpretations of what the sample means might be suggesting about the population parameters (μ's), 2) manually construct orthogonal decision sets to express those possibilities, 3) obtain the implied μ's and the statistics that convey the degree of fit between the sample and implied means for each such set, and then 4) establish or risk her/his scientific reputation by making very precise and specific claims

[6] See section three of Rodger (1974) for an important discussion of the problems associated with analyzing a factorially designed experiment with the usual factorial form of analysis. One important consequence of the increased power to detect non-zero treatment effects that Rodger's method affords is that statistically "significant" interactions will frequently be found when researchers use it. Rodger comments as follows: "In such instances the investigator is faced with the difficult problem of interpreting these interactions. This difficulty arises not from the use of $F[0.05]; \upsilon_1, \upsilon_2$ but from the use of the factorial model (24). It is foolish to use a model with parts which are difficult to interpret. The problem of interpretation is usually simplified if a one-way ANOVA (23) is used and there will be no decrease in β in such analyses using $F[\mathrm{E}\alpha]; \upsilon_1, \upsilon_2$. The researcher is then free to interpret not only contrasts across the μ_{ij} such as (25), (26) and (27), but also simple cross-cell contrasts such as μ_{11} - μ_{22} which are easy to interpret. Simple cross-cell contrasts represent the interactions defined by common sense, though they are not the interactions defined in the factorial model" (p. 195). As recommended in this quoted passage, SPS uses a one-way ANOVA for factorially designed experiments, but it permits the standard two-digit (three-digit) notation for identifying the means in a two-way (three-way), repeated-measures or between-subjects factorial design to be retained (e.g., μ_{23} instead of μ_6).

about the population μ's based on the implied true μ's of the decision set that was finally adopted.

SPS does not require, but it does permit, this usual connection between these two components of the Rodgerian task (construction and selection of a scientifically meaningful decision set) to be broken. Like most other things that can be learned, constructing mutually orthogonal decision sets is experienced as being difficult mainly by those unfamiliar with the task. This particular burden (if it is one), though, can be reduced or completely obviated by taking advantage of one of the several forms of contrast-generation assistance that SPS provides. By offering this help, SPS makes using Rodger's method accessible to every researcher, regardless of his/her skill at constructing mutually orthogonal contrasts. It can probably help many newcomers to the orthogonal-contrast construction task get better at doing this themselves, or, alternatively, it can certainly enable them to avoid learning it altogether.

SPS does not impose any type of contrast-generation assistance on users of the program, but it can make the task of constructing sets of J-1 mutually orthogonal contrasts easier, create such sets based on user-supplied information, or completely automate the process and produce atheoretical decision sets that are composed exclusively of "simple" contrasts (where all of the contrast coefficients are integers, none of which exceeds ±J-1). Whenever the overall ANOVA variance ratio (F_{m}) for a particular study is equal to or greater than F[Eα], anyone using the SPS version of Rodger's method is expected to construct decision sets by themselves, and/or choose some SPS-generated sets, and consider their scientific merit. It is essential that each researcher who uses Rodger's method understand what the individual contrasts in the adopted decision set are asserting, and that they collectively be scientifically meaningful. Ultimately, however, and regardless of how it was conceived and brought into existence, exactly one mutually orthogonal (or merely linearly independent) decision set will be adopted. Those J-1 contrasts, and their assigned g values, are the statistical decisions that the researcher has chosen to make, and the population parameters that those decisions logically imply are the outcome of that study which will be interpreted for the consumers of this research.

Concluding Statement

My Ph.D. dissertation supervisor long ago advised that researchers should make scientifically-informed decisions after carefully considering the data collected in their study, and take their chances on being wrong. If you make J-1 decisions within the context of Rodger's method, you will get, absolutely free, an explicitly stated set of theoretical population parameters (the implied true μ's) that your particular decisions about population μ's mathematically entail. And that's not all. The first part of the judgment expressed by Williams et al. (1992) that is quoted in footnote 1 would almost certainly have been true if it had been written nineteen years before it was, and it remains just as likely to be a true proposition nineteen years later: "Rodger's method ... is the most powerful post hoc method available for detecting true differences among groups" (p. 43).

Rodger's method would seem a nearly irresistible offer, so why have there been so few takers? A more important question concerns its fate. Will Rodger's method continue to be used by only a few researchers, become extinct, or supplant most or all of the currently popular *post hoc* procedures following ANOVA? This article and the SPS computer program constitute an attempted intervention in the competition for dominance and survival that occurs among ideas. My hope is that the power and other virtues of Rodger's method will become much more widely known and that, as a consequence, it will become commonly used. For this to happen, though, some influential statisticians, as well as the powers-that-be at the commercial statistical software companies, are going to have to give Rodger's method some belated attention. In the meantime, anyone who would like to obtain a fairly easy-to-use, Windows-based computer program that implements this very impressive method for making *post hoc* statistical decisions can download one at: http://sites.google.com/site/SPSprogram.

Perhaps the availability of the SPS computer program will help Rodger's method begin to achieve the recognition and use it long-ago deserved. Better ideas and the 'mousetraps' they are instantiated in, ought, eventually, to come to the fore.

Author Note: I want to express my deep appreciation of Bob Rodger, and to thank him for his extensive and patient help over the past several years in sharpening both my understanding of his statistical method and this article. Correspondence concerning this article should be addressed to Mark Roberts, and sent via email to: SPSprogram@gmail.com.

References

Delamater, A. R., Campese, V., & Westbrook, R. F. (2009). Renewal and spontaneous recovery, but not latent inhibition, are mediated by gamma-aminobutyric acid in appetitive conditioning. *Journal of Experimental Psychology: Animal Behavior Processes*, *35*, 224–237. doi:10.1037/a0013293

Rodger, R. S. (1969). Linear hypotheses in 2xa frequency tables. *British Journal of Mathematical and Statistical Psychology*, *22*, 29-48.

Rodger, R. S. (1974). Multiple contrasts, factors, error rate and power. *British Journal of Mathematical and Statistical Psychology*, *27*, 179-198.

Rodger, R. S. (1975a). The number of non-zero, *post hoc* contrasts from ANOVA and error-rate I. *British Journal of Mathematical and Statistical Psychology*, *28*, 71-78.

Rodger, R. S. (1975b). Setting rejection rate for contrasts selected *post hoc* when some nulls are false. *British Journal of Mathematical and Statistical Psychology*, *28*, 214-232.

Rodger, R. S. (1978). Two-stage sampling to set sample size for *post hoc* tests in ANOVA with decision-based error rates. *British Journal of Mathematical and Statistical Psychology*, *31*, 153-178.

Urcuioli, P. J. (2008). Associative symmetry, antisymmetry, and a theory of pigeons' equivalence-class formation. *Journal of the Experimental Analysis of Behavior*, *90*, 257–282. doi:10.1901/jeab.2008.90-257

Williams, D. A., Frame, K. A., & LoLordo, V. M. (1992). Discrete signals for the unconditioned stimulus fail to overshadow contextual or temporal conditioning. *Journal of Experimental Psychology: Animal Behavior Processes*, *18*, 41-55.