Journal of Methods and Measurement in the Social Sciences,
Vol.4, No 1, 20-47, 2013

Comparison of Power for Multiple Comparison Procedures

Robert S. Rodger
Dalhousie University, Nova Scotia

Mark Roberts
British Columbia

The number of methods for evaluating, and possibly making statistical decisions about, null contrasts - or their small sub-set, multiple comparisons - has grown extensively since the early 1950s. That demonstrates how important the subject is, but most of the growth consists of modest variations of the early methods. This paper examines nine fairly basic procedures, six of which are methods designed to evaluate contrasts chosen *post hoc*, i.e., after an examination of the test data. Three of these use experimentwise or familywise type 1 error rates (Scheffé 1953, Tukey 1953, Newman-Keuls, 1939 and 1952), two use decision-based type 1 error rates (Duncan 1951 and Rodger 1975a) and one (Fisher's LSD 1935) uses a mixture of the two type 1 error rate definitions. The other three methods examined are for evaluating, and possibly deciding about, a limited number of null contrasts that have been chosen independently of the sample data - preferably before the data are collected. One of these (planned *t*-tests) uses decision-based type 1 error rates and the other two (one based on Bonferroni's Inequality 1936, and the other Dunnett's 1964 Many-One procedure) use a familywise type 1 error rate. The use of these different type 1 error rate definitions[A] creates quite large discrepancies in the capacities of the methods to detect true non-zero effects in the contrasts being evaluated. This article describes those discrepancies in power and, especially, how they are exacerbated by increases in the size of an investigation (i.e., an increase in J, the number of samples being examined). It is also true that the capacity of a multiple contrast procedure to 'unpick' 'true' differences from the sample data is influenced by the type of contrast the procedure permits. For example, multiple range procedures (such as that of Newman-Keuls and that of Duncan) permit only comparisons (i.e., two-group differences) and that greatly limits their discriminating capacity (which is not, technically speaking, their power). Many methods (those of Scheffé, Tukey's HSD, Newman-Keuls, Fisher's LSD, Bonferroni and Dunnett) place their emphasis on one particular question, "Are there any differences at all among the groups?" Some other procedures concentrate on individual contrasts (i.e., those of Duncan, Rodger and Planned Contrasts); so are more concerned with how many false null contrasts the method can detect. This results in two basically different definitions of detection capacity. Finally, there is a categorical difference between what *post hoc* methods and those evaluating pre-planned contrasts can find. The success of the latter depends on how wisely (or honestly well informed) the user has been in planning the limited number of statistically revealing contrasts to test. That can greatly affect the method's discriminating success, but it is often not included in power evaluations. These matters are elaborated upon as they arise in the exposition below.

Keywords: Multiple comparisons, *post hoc* contrasts, decision-based error rate, power loss

Contrasts and Alternatives

When J random samples of observations are examined, the purpose is very often to find out whether there are differences between them (especially in their averages or means, m_j) that are larger than can reasonably be attributed to random sampling variation or to random assignment of 'subjects' to 'treatments'. Examining differences in averages (called 'comparisons') is a popular way to judge these things, but such simple functions have their limits; so differences between a single group average and the average of K other groups, or between the averages of two sets of groups, can often be more revealing. All these procedures are captured in the theory of contrasts (across means), in both their null and alternative forms. Here we start with the general forms, then move to specific examples.

Generally, a null contrast across the true means (μ_j) of J populations has the form:

$$c_1\mu_1 + c_2\mu_2 + \ldots + c_J\mu_J = 0 \qquad \{1\}$$

in which the c_j are real numbers, not all zero, which sum to zero. They are applied to the sample means (m_j), and have usually been selected by the investigator to reveal what it is believed the sample means say about the relations among the true μ_j.

When {1} is not true, what is true is the alternative:

$$c_1\mu_1 + c_2\mu_2 + \ldots + c_J\mu_J = \delta = g\sigma\sqrt{(\Sigma c^2_j)} \qquad \{2\}$$

Here δ is the linear noncentrality parameter and, if the usual statistical distribution theory is to be used (e.g., the variance-ratio distribution), δ must be expressed in the units of the unknown standard deviation (σ); so the Greek letter has to be there.[1] Since the presence of σ absorbs the scale of measurement used (be it centimetres or inches, kilogrammes or pounds, minutes or seconds, etc.) and $\sqrt{(\Sigma c^2_j)}$ absorbs the scale with which the contrast is expressed [so $(\mu_1+\mu_2)/2 - \mu_3 = g\sigma\sqrt{(1.5)}$ is equivalent to $\mu_1 + \mu_2 - 2\mu_3 = g\sigma\sqrt{(6)}$], the important quantity is g, which is a scale-free parameter. It was created by Rodger (1975b, p. 215) and is not the same thing as g by Hedges (1981). The quantity g is obviously a very important parameter, and is further discussed below, especially in the section 'Choice of g'.

[1] A two-stage procedure created by Stein (1945), with tables provided by Rodger (1976, 1978), can be used for numeric alternatives (without σ).

The best known contrast is the comparison:

$$\mu_1 - \mu_2 = 0 \qquad \{3\}$$

but others of equal, or greater, importance include:

$$(\mu_1+\mu_2)/2 - \mu_3 \equiv \mu_1 + \mu_2 - 2\mu_3 = 0 \qquad \{4\}$$

which compares the average of the first two population means with the third mean, and:

$$(\mu_1+\mu_2)/2 - (\mu_3+\mu_4)/2 \equiv \mu_1 + \mu_2 - \mu_3 - \mu_4 = 0 \qquad \{5\}$$

which compares the average of the first two population means with the average of the second two; and that type of arrangement can go on and on. Contrasts of these types are really essential if mean differences are to be detected efficiently.

Noncentrality and Power: The basic theory of statistical power is due to Neyman and Pearson (1928a, 1928b, 1933a, 1933b) but, until the 1990's, its practical application had been largely ignored in the business of designing statistical investigations. Since then an increasing number of papers have been published on how to calculate a sufficiently large sample size (N) in order to ensure a reasonable probability (power β) of detecting a specified, true, non-zero effect. In spite of all that, research (using statistical methods) reported in journal articles typically have had sample sizes (N) that would yield rather low probabilities (β) of detecting even moderate-sized, true, non-zero effects. (In this paper, detecting where true, non-zero effects exist is taken to be the main purpose of power.) Various explanations have been put forward as to why 'underpower' continues in published papers, and a number of those are discussed by Morrison (2004), who also cites a number of papers that report the details of this 'underpower' in various sub-fields. Unfortunately, Morrison cites none of Rodger's papers on detection rate, though this current paper shows that the Rodger method is particularly simple, practical and effective.

If the variance-ratio distribution is to be used in the analysis (e.g., rather than the Studentized range distribution), then it is a quadratic noncentrality parameter that is required in that distribution to compute power. For the hth contrast that quantity can be written as:

$$\Delta_h = N\delta^2_h/(\sigma^2\Sigma_j c^2_{hj}) = Ng^2_h\sigma^2(\Sigma_j c^2_{hj})/(\sigma^2\Sigma_j c^2_{hj}) = Ng^2_h \qquad \{6\}$$

(see {2} above) and that is part of the overall, quadratic, noncentrality parameter for analysis of variance (anova) to evaluate the classical null hypothesis:

H_0: $\mu_1 = \mu_2 = . . . = \mu_J$. {7}

Clearly, {6} shows another very important property of g, i.e., its simple relationship to the F noncentrality parameter Δ. Also, H_0 at {7} is true when any set of H = J-1, linearly independent[2], null contrasts are true. An important example is any set of mutually orthogonal (i.e., uncorrelated) null contrasts, such as the comparison-based set:

$\mu_1 - \mu_2 = 0$ {8}
$\mu_3 - \mu_4 = 0$ {9}
$\mu_5 - \mu_6 = 0$ {10}
$\mu_1 + \mu_2 - \mu_3 - \mu_4 = 0$ {11}
$\mu_1 + \mu_2 + \mu_3 + \mu_4 - 2\mu_5 - 2\mu_6 = 0$ {12}

Sets of contrasts are often represented in a matrix of their contrast coefficients (c_{hj}), such as the above H = 5, shown in Table 1:

Table 1
Mutually Orthogonal Contrast Coefficients (c_{hj})

h	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	$\Sigma_j c^2_{hj}$
1	1	-1	0	0	0	0	2
2	0	0	1	-1	0	0	2
3	0	0	0	0	1	-1	2
4	1	1	-1	-1	0	0	4
5	1	1	1	1	-2	-2	12

The null hypothesis at {7} is commonly evaluated by the statistic F_m in anova, i.e.:

[2] Linear independence is essential to avoid repetition, and especially contradiction. If one makes contradictory assertions, all those assertions become worthless. For example, to assert 'because the statistics say so' that μ_1-μ_2=0, μ_2-μ_3=0 and μ_1-μ_3<0, is a contradiction, no matter what unthinking statistics one used on the m_j. It was a 'common notion' of Euclid (who lived around 300BC) that two things (e.g., μ_1 and μ_3) that are equal to the same thing (e.g., μ_2), are equal to one another. The statements μ_1-μ_2=0, μ_2-μ_3=0 and μ_1-μ_3=0 are repetitious because any two of these implies the third. In each of these two illustrations, the elements of the three statements constitute a linear equation, since A:μ_1-μ_2=0, B:μ_2-μ_3=0, C:μ_1-μ_3=0 are linearly related by A + B = C; equivalent to ($\mu_1 - \mu_2$) + ($\mu_2 - \mu_3$) = $\mu_1 - \mu_2 + \mu_2 - \mu_3 = \mu_1 - \mu_3$. It seems strange to have to spell this out, but there continue to be scientific papers that make contradictory (linearly dependent) assertions, or something a little more vague but effectively equivalent to that!

$$F_m = N\Sigma(m_j - m.)^2/(\nu_1 s^2) \qquad \{13\}$$

in which m. is the mean of the sample means (m_j), ν_1 = J-1 is the numerator degrees of freedom for F_m, and s^2 is the error variance (based on ν_2 = J(N-1) degrees of freedom).

When the null hypothesis at {7} is not true, then the distribution of F_m is the variance ratio distribution with degrees of freedom ν_1 and ν_2, but also with quadratic, noncentrality parameter:

$$\Delta_m = N\Sigma(\mu_j - \mu.)^2/\sigma^2 \qquad \{14\}$$

in which $\mu.$ is the mean of the μ_j and σ^2 is the true variance. It should now be clearer than ever where the σ in {6} came from. And one of the many beauties of the anova system is the algebraic similarity between Δ_m at {14} and F_m at {13}. Furthermore, when {7} is true, Δ_m = 0. But that is by no means all, because for any set of (J-1) mutually orthogonal contrasts, each with a true value g_h, then:

$$\Delta_m = N\Sigma g^2{}_h \qquad \{15\}$$

A similar relation exists for any J-1 linearly independent contrasts (that need not be mutually orthogonal). That procedure involves a matrix product and a matrix inverse because non-orthogonal contrasts share (i.e., overlap) the variation among the m_j (and among the μ_j).

Equal N Used: So far, all the formulae have used a sample size (N) which is constant from sample to sample. That constant-N rule will continue in this article. There are formulae for unequal sample sizes (N_j in sample j) but they are somewhat more complicated than the constant-N forms; so not so easy to follow. Also, the use of unequal N_j raises the risk of weakening the validity of the procedures if the population true variances ($\sigma^2{}_j$) happen to be unequal; so should not be encouraged when analyzing means.

The Post Hoc Methods

There are basically two approaches to disentangling inter-sample (actually inter-population) differences. If one has enough detailed information about where true inter-population mean differences lie, one can **plan** to test a set of J-1 linearly independent null contrasts before the random sample data are collected, choosing sample size (N) to yield a reasonably high probability (β) of rejecting each of the false null contrasts (true nulls will, hopefully, be retained), then test each contrast with a type 1 error rate α (e.g., using a two-tailed *t*-test, or its equivalent $t^2 = F\alpha;1,\nu_2$).

The same procedure applies if one's scientific, theoretical understanding of the research topic provides a very clear idea of where the true differences among the μ_j should lie. In that case, one should (prior to obtaining the data) plan J-1 contrasts for testing that share out the sizes of effects more or less equally among the potentially false nulls. All of that sounds so unrealistic that it may apply only very rarely. Therefore, except for two-group studies, research that uses '**planned contrasts**' is likely to be treated with suspicion, especially if the fit of the tests to the sample data is rather close!

The other approach is to choose the J-1 contrasts for decision making in the light of how the sample data turn out. That is the ***post hoc*** strategy, and one can pre-calculate the size (N) one's samples need to be to detect contrast effects of pre-specified size, with expected detection rate Eβ. This approach can be criticized on the ground that it does not follow the hypothetico-deductive method of science (i.e., state a theory, deduce its observable consequences, collect appropriate data, and check that they are consistent with the consequences deduced from the theory). But if error rate and power can be properly controlled, the *post hoc* strategy has much to recommend it. We often have a number of different, theoretical conceptions, and sometimes need data to indicate where differences do and do not lie; so theory deduction and observed confirmation are rather idealistic. Once theoretical conceptions become clearer, it may then be possible to check them with a few, carefully-chosen planned contrasts! Apart from satisfying statistical criteria, an essential requirement is that whatever decisions the *post hoc* procedure yields should make scientific sense.

Of the six different methods (examined here) that have been used to evaluate contrasts *post hoc*, i.e., after the data have been collected, examined and given a preliminary analysis, four use experimentwise or familywise type 1 error rates. Usually such preliminary analysis is an analysis of variance (anova), but other procedures include the analysis of proportions and of ranked data. The *post hoc* methods reported on here that use experimentwise error rate are those of Scheffé and of Tukey. Newman and Keuls use a familywise error rate, and Fisher's LSD uses a mixture of experimentwise and decision-based type 1 error rates.

Scheffé: This method says whether any contrast (h) across the sample means (m_j) is consistent with the value of the true-means (μ_j) version being zero; maybe even deciding against that[B] if:

$$F_h = N(\Sigma_j c_{hj} m_j)^2/(\nu_1 s^2 \Sigma_j c^2_{hj}) \geq F\alpha;\nu_1,\nu_2 \qquad \{16\}$$

in which α is an experimentwise type 1 error rate, such as 0.05 or 0.01. This is wonderfully consistent with the test of H_o at {7}, by rejecting that

H_o if the overall F_m at {13} is $\geq F\alpha;\nu_1,\nu_2$. If that overall test fails, no contrast in the data will be able to satisfy the {16} criterion.

Tukey: Tukey defined his “Honestly Significant Difference” (HSD) as:

$$|\Sigma c_j m_j| \geq q\alpha;J,\nu_2\ 0.5\Sigma|c_j|\sqrt{(s^2/N)} \qquad \{17\}$$

where $q\alpha;J,\nu_2$ is the critical Studentized range statistic, for J groups, with ν_2 error degrees of freedom, and α is an experimentwise error rate such as 0.05 or 0.01. If the largest m_j minus the smallest (i.e., the range of the m_j) is not large enough, no other contrast across the m_j will be large enough to meet the criterion.

Newman-Keuls Multiple Range: This (NKMR) procedure by these two authors is for comparisons only and, like Tukey, uses a Studentized range statistic. Since only comparisons can be evaluated, then $0.5\ \Sigma|c_j| = 1.0$ for all cases, and the NKMR starts with the largest mean difference (compared to $q\alpha;J,\nu_2$) and if that comparison 'makes that grade', the NKMR procedure then looks at the next largest sub-range (i.e., largest mean minus second-smallest, and second-largest minus smallest), but evaluated against a sub-range of K = J-1 means, and so on. The formula (compare with {17}) is:

$$|m_i - m_j| \geq q\alpha;K,\nu_2 \sqrt{(s^2/N)} \qquad \{18\}$$

Here α is one of the conventional probabilities (such as 0.05 or 0.01) but it is not, strictly speaking, an experimentwise type 1 error rate (except when K = J). More generally, α is a familywise error rate (for various sub-range families).

Fisher's LSD: This fourth method begins by evaluating H_o at {7} by comparing F_m at {13} against $F\alpha;\nu_1,\nu_2$ as in a traditional anova (much as Scheffé's procedure might do). That is using an experimentwise type 1 error rate. Only if $F_m > F\alpha;\nu_1,\nu_2$ will the LSD method proceed to evaluate any and all contrasts of interest by *t*-tests. Since the square of a *t* value is an *F* value with $\nu_1 = 1$, the procedure is to see whether any contrast (h) across the sample means (m_j) is consistent with the value of the true-means (μ_j) contrast being zero; maybe even deciding against that, if:

$$F_h = N(\Sigma_j c_{hj} m_j)^2/(s^2 \Sigma_j c^2_{hj}) \geq F\alpha;1,\nu_2 \qquad \{19\}$$

This is a formula like Scheffé's at {16}, only here $\nu_1 = 1$ not (J-1). The α used here is a conventional one (such as 0.05 or 0.01); so this is (supposedly) a decision-based type 1 error rate. But if H_o at {7} is true (hence all null contrasts are true), the effective type 1 error rate is notably larger than the supposed α within those experiments in which H_0 has been rejected in error.

When LSD Rejects a True H_0: The F_m at {13} had $\nu_1 = (J-1)$ as a divisor; so to compare what it reports against the $F\alpha;1,\nu_2$ at {19}, we must multiply F_m by (J-1). Note that if $(J-1)F_m > F\alpha;1,\nu_2$ (the square of the critical *t* value), then we can ALWAYS find J-1 linearly independent, null contrasts to reject (by $F\alpha;1,\nu_2$) in the sample data. But such contrasts are not necessarily very simple. For example, contrasts such as:

$h = 1, \mu_1 - \mu_2 = 0$ {20}
$h = 2, 4\mu_1 - 3\mu_2 - \mu_3 = 0$ {21}
$h = 3, 7\mu_1 - 4\mu_2 - 6\mu_3 + 3\mu_4 = 0,$ {22}

though not very simple, are much simpler than others that might be necessary to squeeze into the 'rejection space' between $(J-1)F_m$ and $F\alpha;1,\nu_2$. The three contrasts above are intercorrelated $r_{12} = 0.97$, $r_{13} = 0.74$ and $r_{23} = 0.86$ but, although each pair of contrasts shares between them a good deal of the variation among the m_j, they are linearly independent of one another. That 'sharing' of the variation could be much, much closer if desired and, in that way, highly correlated null contrasts could all be rejected even if the 'space' between $(J-1)F_m$ and the rejection criterion ($F\alpha;1,\nu_2$) is quite small. Of course, closely correlated contrasts do not tell us much more (about the μ_j) than fewer more widely separated contrasts. At the limit, the most separated are orthogonal contrasts, and each of these reveals information about the true μ_j that is more or less independent of the others.

To pursue further the notion of mutually orthogonal contrasts only, consider the rule (analogous to {15}) that says F_m is the sum of F_h for orthogonal contrasts. It follows that, when α = 0.05 and J = 4, N = 6:

$$(J-1)F\alpha;J-1,\nu_2/F\alpha;1,\nu_2 = 3F0.05;3,20/F0.05;1,20 \quad \{23\}$$
$$= 3\times3.098/4.351 = [2.1] = 2,$$

which tells us that LSD can reject at least 2 out of J-1 = 3 mutually orthogonal null contrasts when H_0 at {7} can be rejected (i.e., a 2/3 type 1 error rate when H_0 was rejected in error). When J is larger, the number of erroneous rejections is worse. When J = 12, N = 6, the {23} ratio becomes:

$$(J-1)F\alpha;J-1,\nu_2/F\alpha;1,\nu_2 = 11F0.05;11,60/F0.05;1,60 \quad \{24\}$$
$$= 11\times1.952/4.001 = [5.4] = 5$$

and when J = 24, N = 6:

$$(J\text{-}1)F_{\alpha;J\text{-}1,\nu_2}/F_{\alpha;1,\nu_2} = 23F_{0.05;23,120}/F_{0.05;1,120} \qquad \{25\}$$
$$= 23\times 1.620/3.920 = [9.5] = 9$$

Though the number of rejections has increased, the ratios of rejections/(J-1) have decreased. But those numbers of erroneous null rejections are minima, because the observed F_m, against which the numerators ($F_{\alpha;J\text{-}1,\nu_2}$) are compared, are likely to be larger.

Of course, the type 1 error rate is zero for all the contrasts in each of the experiments in which H_o at {7} has been accepted correctly (assuming the investigator believes in accepting nulls)! But the general picture is one of infrequent bursts of many errors, and long sessions with no error at all. A rather weird way to work, according to the authors, who prefer to sprinkle errors, little by little, as they proceed! The effect of all this on detecting false nulls is discussed below, following Table 4.

Fisher's LSD is a way of "keeping most of your 'rotten' eggs in few baskets." But as is shown in Table 2 for the Scheffé procedure, your chance of recovering the real, healthy eggs diminishes dramatically as the basket grows in size!

The LSD acronym for Least Significant Difference is amusing because, for centuries, the British used that acronym to refer to their pre-decimal currency L (librae, £, pounds), S (solidi, shillings), and D (dinarii, pence). But, as shown above and following Table 4 below, there is not much 'real' money in Fisher's LSD!

The other two *post hoc* procedures are due to Duncan (1951, 1952, 1955) and to Rodger (1967a, 1967b, 1974, 1975a, 1975b). These both use decision-based type 1 error rates.

Duncan's Multiple Range: The rationale Duncan gave for his multiple range procedure (DMR) is not easy to follow, but one that seems to fit the philosophy is Rodger's (1967a) original concept. That is, suppose a researcher only ever studies two samples at a time, analyses the mean difference, then publishes the result. If we select K of her/his reports, in which the K sample-pairs are independent of one another then, if $\mu_1\text{-}\mu_2 = 0$ had been true for every one of those K reports, and if type 1 error rate α had always been used, the probability that one or more of those K nulls had been rejected in error would be $\gamma = 1\text{-}(1\text{-}\alpha)^K$. That is the pronouncement of Bernoulli's Binomial Theorem! If $\alpha = 0.05$ then $\gamma = 0.19, 0.34, 0.46, 0.71$ when K = 4, 8, 12, 24. Those are embarrassingly high probabilities of error, but the researcher's procedure is beyond reproach (except maybe having better efficiency by studying more than just two samples at time). If there's a cause for concern, it's more likely to be the standard used for judgment (i.e., committing one or more errors).

That being so, surely a researcher who studies J = 9 groups at a time (and makes J-1 = 8 decisions) should be allowed to tolerate the probability of one or more errors among the eight to be 0.34.

Duncan's DMR is a step-down method like the Newman-Keuls multiple range procedure, but with a different (decision-based) familywise type 1 error rate (if that is not too confusing a concept). Thus for Duncan, a comparison across the sample means (m_j) is not consistent with a zero value of the true-means (μ_j) comparison if:

$$|m_i - m_j| \geq q\gamma;K,\nu_2 \sqrt{(s^2/N)} \qquad \{26\}$$

where $\gamma = 1-(1-\alpha)^{K-1}$, and K is the step-down sub-range of the means compared. [Note how {26} is similar to {18}.] When a comparison fails the {26} criterion, no other comparison inside that failed range can be allowed to pass its {26} test, no matter what the data in that comparison say.

Rodger: The original proposal (1967a) was similar to Duncan's, though Rodger did not know of Duncan's work at the time (i.e., Rodger used $\gamma = 1-(1-\alpha)^{\nu_1}$ to evaluate any and all contrasts across the means m_j). But the variance ratio distribution was used (not the Studentized range), there was no restriction on contrast forms (i.e., the method was not just for comparisons), and the procedure was to find H = J-1 linearly independent contrasts across the J values of m_j (preferably H = J-1 mutually orthogonal contrasts) among which r nulls - given by {27} - would be rejected and ν_1-r accepted. By 1966, even before that first 1967 paper had appeared in print, Rodger realized that control of the average rate of null rejection (i.e., the expectation of r/ν_1) would be a far better quantity to control (than the probability of rejecting one or more nulls in error); so he published tables (1975a) of the new criterion $F[E\alpha];\nu_1,\nu_2$ rather than his original $F\gamma;\nu_1,\nu_2$. Rodger's procedure is, first compute:

$$r = [F_m/F[E\alpha];\nu_1,\nu_2] \leq \nu_1 \qquad \{27\}$$

in which the outer [ ] indicate that any fraction must be deleted, and the ≤ sign says r cannot be allowed to exceed ν_1, because no more than ν_1 linearly independent contrasts are mathematically possible across J-1 = ν_1 means. In order to fit the sample data better, mutually orthogonal contrasts are preferred (r of these are always possible, but some of them may be just too hard to interpret scientifically). If the rule at {27} is followed, the expected (average) rate of rejection of true null contrasts will be Eα when H_0 at {7} is true.

Rodger (1975b) also published tables of the parameters $\Delta[E\beta];\nu_1,\nu_2$ which allow one to calculate the sample size (N) necessary to give one's research project the probability $E\beta$ of detecting null contrasts that are false by an amount ±g (or, more precisely, to bring the expectation of r/ν_1 close to $E\beta$, if the variation among the μ_j – as given by {14} – is at least $\nu_1 N g^2$).

Before an investigation starts, the investigator should work out (from the study of previous research on the topic) the size of the treatment effect (g^2) he/she would like to detect (if it exists), the rate ($E\beta$, e.g., 0.95) at which he/she wishes detection to occur, then calculate:

$$N \geq \Delta[E\beta];\nu_1,\nu_2/g^2 \qquad \{28\}$$

beginning by using $\nu_2 = \infty$.

The procedure is much simpler than it sounds, it is illustrated by examples in Rodger's cited papers, and more information can be found, including a worked example, at:

http://en.wikiversity.org/wiki/Rodger%27s_Method

The Simple, Powerful Statistics (SPS) computer program carries out various Rodgerian statistical procedures, including sample size calculations as well as non-parametric analyses of proportions (as shown in Rodger 1969) and ranks. It also reports the values (in σ units) of the parameters (e.g., the $\mu_j - \mu.$) implied by the statistical decisions made. SPS is a free, Windows-based program that can be downloaded at:
http://sites.google.com/site/SPSprogram

An article describing both Rodger's method and the SPS program, which makes using it accessible to researchers, was published by the SPS creator Roberts (2011).

Illustrative Power Comparisons

To illustrate the differences in power between the methods, suppose we have normally distributed variates which have true means $\mu_1 = 70$, $\mu_2 = 50$, $\mu_3 = \mu_4 = . . . = \mu_J = 60$, all with the common variance $\sigma^2 = 100$. The difference:

$$\mu_1 - \mu_2 = 20 = g\sigma\sqrt{\Sigma c^2_j} = 1.414\sigma\sqrt{2} \qquad \{29\}$$

has a very large g (1.414) - and more will be said about that below (see the section 'Choice of g') - but it allows us to use small samples to show reasonable power (β) when J is small.

Illustration Data: We will draw random samples of N = 6 from J of these populations, use α = 0.05 (or Eα = 0.05) everywhere, and analyze the mean differences by (1) Tukey's (1953) HSD procedure for contrast evaluation. Procedure (2) will be Duncan's (1951) multiple range method (DMR, for comparisons only). Both of these methods use the Studentized range distribution, but with different definitions of type 1 error rate. Technique (3) will be Scheffé's (1953) method. Technique (4) will be Rodger's (1975b) method. Both of those methods use the variance ratio distribution, with $\alpha = 0.05$ for Scheffé and $E\alpha = 0.05$ for Rodger. Because we will always use the first J groups of $j = 1, 2, \ldots$, the overall, quadratic noncentrality parameter for the variance ratio distributions, will always be:

$$\Delta_m = N\Sigma(\mu_j - \mu.)^2/\sigma^2 \qquad \{30\}$$
$$= 6(10^2+(-10)^2+0+\ldots+0)/100 = 12.000$$

no matter what value of $J \geq 2$ is chosen.

The Studentized range distribution does not have a noncentrality parameter which, in itself, is a serious limitation. The Studentized range distribution uses the standardized true means to find the power, i.e.:

$$(\mu_j - \mu.)/\sqrt{(\sigma^2/N)} = (70 - 60)/\sqrt{(100/6)} \qquad \{31\}$$
$$= 10/4.082 = 2.449$$

for $\mu_1 - \mu.$; $\mu_2 - \mu.$ is -2.449 and all the other $\mu_j - \mu.$ will be zero.

The pattern of μ_j used here is very important because matters are not comparable if Δ_m is increased as J is increased - as has been allowed in some research on power - or if the value of the range of the standardized true means is increased as J is increased.

Detecting a False H_0: Table 2 shows the results of numeric integrations, in which βT is the power for Tukey's (and the NKMR) method, βD the power of Duncan's (DMR) procedure (both using the Studentized range distribution), βS the power for Scheffé's technique, and βR the power for Rodger's method (both using the variance ratio distribution). All of those β values are the probabilities of rejecting H_0 at {7}, or its equivalent, although Rodger does not treat that H_0 as an hypothesis of primary interest. β should be the same for all four methods when J = 2; the differences are due to computer rounding.

Both Tukey's (and Newman-Keuls multiple range, NKMR) method and Scheffé's technique lose power quite dramatically as J increases: Tukey by more than 30% when J = 12, and by more than 40% when J = 24, Scheffé by more than 38% when J = 12 and by almost 55% when J = 24. Tukey would need N = 10 to maintain $\beta \geq 0.8763$ and Scheffé would need N = 11 to maintain $\beta \geq 0.8764$ when J = 12 (both through increasing

Table 2

Powers (β) of Rejecting H_0 by Various Methods

	Num Gps J =	2	4	6	8	10	12	24
Method	ErrorDF ν_2 =	10	20	30	40	50	60	120
Tukey &	$q_{0.05;J,\nu_2}$	3.151	3.958	4.302	4.521	4.680	4.808	5.266
NKMR	βT	.8763	.7643	.7051	.6641	.6331	.6072	.5098
Duncan	$q(0.05);J,\nu_2$	3.151	3.190	3.250	3.300	3.340	3.374	3.498
DMR	βD	.8763	.9015	.9144	.9241	.9322	.9392	.9674
	Num DF ν_1 =	1	3	5	7	9	11	23
Scheffé	$F_{0.05;\nu_1,\nu_2}$	4.965	3.098	2.534	2.249	2.073	1.952	1.620
	βS	.8764	.7546	.6787	.6214	.5755	.5372	.3949
Rodger	$F[0.05];\nu_1,\nu_2$	4.965	2.126	1.499	1.226	1.068	0.961	0.620
	βR	.8764	.8875	.9059	.9215	.9355	.9483	.9946

noncentrality and ν_2). One must increase N in these notable ways, as J is increased, to maintain decent power; otherwise true non-zero detection capacity will be drastically reduced. That is neither a characteristic of nature nor of mathematics, but an artifact of the choice of 'experimentwise' error rates (controlling the rate at which H_0 at {7} will be rejected in error at a conventional value of α) rather than decision-based error rates. [C]

Happily, both Duncan's (DMR) and Rodger's methods not only hold their 'false H_0' detection capacities as J increases, they actually improve them somewhat: just over 7% for Duncan, and just over 8% for Rodger when J goes from 2 to 12.

Unscrambling the μ_j From the m_j

Duncan: Unhappily, the fact that Duncan's procedure is a multiple range method limits its capacity to unravel the likely differences among the μ_j that the m_j indicate. For example, suppose we are using J = 4 and our sample values turn out to be $m_1 = 70$, $m_2 = 50$, $m_3 = 59$, $m_4 = 57$ with $s^2 = 180$. Duncan's procedure would first examine $m_1 - m_2 = 20$ (the largest difference first), using $q(0.05);4,20 = 3.190$, where (p) is used rather than the code in $\gamma = 1\text{-}(1\text{-}p)^3$ to make matters clearer. That would reject:

$$\mu_1 - \mu_2 = 0 \qquad \{32\}$$

because:

$$m_1 - m_2 = 70 - 50 = 20 > q(0.05);4,20\sqrt{(s^2/N)} \qquad \{33\}$$
$$= 3.190\sqrt{(180/6)} = 17.5$$

Moving in from there uses $q(0.05);3,20 = 3.097$; so $m_1 - m_4 = 13$ and $m_3 - m_2 = 9$ need to be at least as large as:

$q(0.05);3,20\ \sqrt{(s^2/N)} = 3.097\sqrt{(180/6)} = 17.0$; {34}

therefore neither $\mu_1 - \mu_4 = 0$ nor $\mu_3 - \mu_2 = 0$ may be rejected. Only the first tested ($\mu_1 - \mu_2 = 0$) may be rejected. But multiple range procedures do not make decision claims (one wonders what rôle α serves for them). These users underline the m_j (not the μ_j) that do not differ by their criterion. Putting the m_j in order of size, our (supposed) data yield:

$$\begin{array}{llll} m_2 & m_4 & m_3 & m_1 \\ \cline{2-4} \\ \cline{1-2} \end{array} \qquad \{35\}$$

What are we to believe about μ_3 and μ_4? Did the investigator believe that the sample evidence indicated that either of those parameters differed from μ_1, or from μ_2? We should not be left guessing what the investigator believes the data indicate.

Rodger: Rodger's method would note that:

$F_m = N\Sigma(m_j - m.)^2/(\nu_1 s^2) = 6\times 206/(3\times 180) = 2.289$ {36}

Hence, Rodger's method (see {27} above) says we may reject:

$r = [F_m/F[0.05];3,20] = [2.289/2.126] = [1.08] = 1$ {37}

null contrast, and the obvious one is $\mu_1 - \mu_2 = 0$ because:

$F_1 = N(m_1 - m_2)^2/(\nu_1 s^2 \Sigma c^2_j) = 6\times 20^2/(3\times 180\times 2)$ {38}
$= 2400/1080 = 2.222 > F[0.05];3,20 = 2.126$

This is not a great achievement, because Duncan's DMR said the same thing. The difference is in what Rodger would do next, i.e., test and decide that:

$\mu_3 - \mu_4 = 0$; $F_2 = 0.022$ {39}

$\mu_1 + \mu_2 - \mu_3 - \mu_4 = 0$; $F_3 = 0.044$ {40}

and for these three orthogonal contrasts:

$F_m = \Sigma F_h = 2.222 + 0.022 + 0.044 = 2.288$ {41} which differs from {36} only by rounding error.

Also, the three decisions for the μ_j tell us that our data support the interpretation:

$\mu_2 < \mu_3 = \mu_4 < \mu_1$ {42}

Accepting Null Contrasts: Note how {42} depends on the two nulls at {39} and {40} being 'accepted'; so those who *never* accept nulls cannot make such a 'logical' connection. Of course, accepting a null contrast is not 'chipped in stone' and other evidence might indicate such 'acceptance' was likely a type 2 error. Furthermore, accepting a null contrast can be construed as acting as if the difference, if any, is negligible in the present state of our knowledge of the topic. Table 3 below provides 'guideposts' on how small is small;[D] at least from a statistical standpoint. We should be sensible about null contrasts, by accepting them when the evidence supports that (we are not 'proving' things with statistics), and make the whole process more rational by designing our investigations to have good power (in particular, good Eβ) to detect effects of reasonable size.

Comparisons Only Limits: Because multiple range methods are restricted to comparisons (i.e., $m_i - m_j$ only), they cannot easily decide that $\mu_2 < \mu_3 < \mu_1$ in the above example (or generally). Although the range statistic may indicate that $\mu_2 < \mu_1$, neither m_2 nor m_1 are different enough from m_3 to claim either $\mu_2 < \mu_3$ or $\mu_3 < \mu_1$ (and similarly for m_4). Testing comparisons only, and no other forms of contrasts (as is the rule for multiple range methods), is a very serious limitation. To reject $\mu_1 - \mu_4 = 0$ (the next sub-range in our illustration) by the DMR would need a much larger sample size, N = 10 (using $q(0.05);3,36 = 3.015$ for the m_j and $s^2 = 180$ in our illustration). To reject $\mu_3 - \mu_2 = 0$ (in that same sub-range) would need an even larger N. That structural limitation exacerbates any reduced power problem that multiple range methods might have.

Scheffé and Tukey: Neither of these methods would reject H_o for our J = 4 example data. Scheffé's method would find:

$$F_m = 6\Sigma(m_j - m.)^2/(3\times 180) \qquad \{43\}$$
$$= 2.289 < F0.05;3,20 = 3.098$$

and Tukey's HSD would find:

$$|m_1 - m_2| = 20 \underline{<} q0.05;4,20\sqrt{(180/6)} \qquad \{44\}$$
$$= 3.958\sqrt{30} = 21.679$$

Hence there would be no further analysis by either of these methods. Table 2 shows that the probability of these methods rejecting H_o correctly were βS = 0.7546 and βT = 0.7643. Luck was not with them, though it was for Duncan's DMR (βD = 0.9015) and Rodger's method (βR = 0.8875), according to Table 2.

Choice of g: It was noted at {29} above that the g value of 1.414 was large, but large g values do sometimes occur in research. Examples include some studies of pigeon learning and perception in which the birds have been so extensively trained that the variation in their behaviour is very slight.

But usually we should be aiming to detect g values of 1.0 or less. As an example of what that means, note that in the anthropometric study of Americans in 2003 to 2006 by McDowell et al. (2008), they found the average adult female and male standing heights to be 162.2 cm (5' 3.8") and 176.3 cm (5' 9.4") with standard deviation 11 cm. That 14.1 cm (= 5.6") difference is:

$$m_m - m_f = 176.3 - 162.2 = 14.1 \approx 0.9\times 11\sqrt{2} \qquad \{45\}$$

so a g ≈ 0.9 or 1.0 is a noticeable difference.

Cohen[E] (1988, 1992) was a strong advocate for choosing N to set power (β), and had recommendations on what he considered to be small, medium and large effects. But all of that was set in the traditional context (of anova and similar procedures) using experimentwise definitions of α and β.

In Rodger's (1975a, 1975b) context of decision-based Eα and Eβ (being rejection rates per decision), a 'large effect' would have g^2 around 1.00, a 'moderate effect' is a g^2 about 0.50, a 'small effect' is a g^2 approximately 0.25 and a 'slight effect' would be g^2 around 0.125. The differences between the two sets of standards are illustrated in Table 3. There it is assumed we are looking for the difference (|g|) in average standing height between two human, adult sub-populations, to be tested by an α = 0.05 *t*-test, and the required N for each of the two random samples (one of males, one of females) to achieve detection probability β ≥ 0.95 is shown. The McDowell data above provided the estimate of σ ≈ 11. It is hoped that the representation of the L, M, S, Sl guideposts as standing height differences in both inches and centimetres will make them more comprehendible.

The adult standing height data should provide a familiar norm, but each investigator will be able to establish norms for the subject matter being studied, assuming a decent estimate of σ^2 is available. The data quoted here from the McDowell et al. (2008) report had 4857 women and 4482 men.

Table 3

N to Make β = 0.95 When α = 0.05 for Height |g|

Standards	$(\mu_m-\mu_f)$cm	$(\mu_m-\mu_f)$in	|g|	g^2	N
McDowell	14.10	5.55	0.906	0.822	17
Cohen L	8.80	3.46	0.566	0.320	42
Cohen M	5.50	2.17	0.354	0.125	105
Cohen S	2.20	0.87	0.141	0.020	650
Rodger L	15.56	6.13	1.000	1.000	14
Rodger M	11.00	4.33	0.707	0.500	27
Rodger S	7.78	3.06	0.500	0.250	53
Rodger Sl	5.50	2.17	0.354	0.125	105

Note: Effect sizes are: L=large, M=medium, S=small, Sl=slight.

Using g^2 = 1: If we had designed our power illustration for Table 2 to have a g ≈ 1.0, that would make $\mu_1 = 70$, $\mu_2 = 56$, $\mu_3 = \mu_4 = \ldots = \mu_J = 63$ and, remembering that $\sigma^2 = 100$:

$$\mu_1 - \mu_2 = 14 \approx 1.0\ \sigma\sqrt{2} \qquad \{46\}$$

To detect this |g| with any of the procedures, setting β ≈ 0.9, would require N = 12. With this double-sized N, if J is increased as in Table 2, the percentage loss of power (β) is similar to that of the smaller N = 6. Thus our revised N = 12 and μ_j make $\Delta_m = 11.760$ and with $F0.05;1,22$ = 4.301, Scheffé's method gives βS = 0.9059 (when J = 2), but with $F0.05;11,132$ = 1.862 (when J = 12), we find βS = 0.5695 (a 37% drop) - and the drop is more as J increases. Similarly, Tukey's (and the NKMR) procedure, using $q0.05;12,132$ = 4.689 and $(\mu_{LG}-\mu.)/\sqrt{(\sigma^2/N)} = (70-63)/\sqrt{(100/12)} = 2.425$, with -2.425 for μ_{SM}, and all other $(\mu_j-\mu.)/\sqrt{(\sigma^2/N)} = 0$, shows a drop from βT = 0.9059 (when J = 2) to 0.6314 (a 30% drop at J = 12), and bigger drops as J increases. Both these methods, of course, have larger $\nu_2 = J(N-1)$ when N = 12. The only *post hoc* methods to hold their power as J increases are Rodger's method (βR) and Duncan's DMR (βD), but Duncan's method has the serious, limiting problem of 'comparisons only' when 'looking' inside the full range.

Unprotected *t*-tests: With this procedure one tests any contrasts at all (after examining the sample data) by a conventional *t*-test (i.e., using the conventional α = 0.05 or 0.01) without any prior check, such as the check on F_m used in Fisher's LSD. Everyone should know that this method inflates the type-1 error rate considerably beyond the 'conventional' α value cited. Nevertheless, this practice continues, and papers explicitly using it are published in what one would think are reputable journals.

Given that $t0.05;\nu2$ is used, what is the actual type-1 error rate when the null contrast is true? One of the simplest ways to examine this is to compare what $t_{0.05;\nu_2}$ would find according to the correctly-appropriate Studentized range distribution (q); for example, for comparisons such as $m_i - m_j$. The critical $q_{crit;J,\nu_2}$ appears in the formula to reject $\mu_i - \mu_j = 0$ when:

$$|m_i - m_j| \geq q_{crit;J,\nu_2} \sqrt{(s^2/N)} \qquad \{47\}$$

and the *t* formula says reject $\mu_i - \mu_j = 0$ when:

$$|m_i - m_j| \geq t_{\alpha;\nu_2} \sqrt{(2s^2/N)} \qquad \{48\}$$

It follows that the q equivalent (say, Q) for *t*, equating the right-hand sides, is:

$$Q = t_{\alpha;\nu_2}\sqrt{2} \qquad \{49\}$$

For example, from Table 2, we see the 'true' equivalence of $F_{0.05;1,10} = 4.965$ (the square of $t_{0.05;10}$) and $q_{0.05;2,10} = 3.151$ as:

$$\sqrt{(2 \times 4.965)} = 3.151 = q_{0.05;2,10} \qquad \{50\}$$

Table 4 shows what type-1 error rate the Studentized range distribution gives for $q_{crit;J,\nu_2} = t_{0.05;\nu_2}\sqrt{2}$. In all cases, N = 6 is used, (the μ_j are all equal) but the m_j are arranged in order of size, smallest on the left to largest on the right. If we have J = 12, then $\nu_2 = 60$, and the tabled $t_{0.05;60} = 2.000$; so in the Studentized range distribution the formula equivalent to that *t* is $Q = t\sqrt{2} = 2.000\sqrt{2} = 2.829$. When that is used to evaluate $m_{12} - m_1$ (the largest observed difference), the type-1 error rate is not the 'assumed' 0.05 but 0.691. That is the integral of the Studentized range distribution for J = 12, $\nu_2 = 60$, and all μ_j equal. With the same J = 12, if we use $t_{0.05;60} = 2.000$ (Q = 2.829) as if in a Newman-Keuls double step-in subrange to evaluate $m_{11} - m_2$ (with K = 10), then the type-1 error rate is 0.602. These results show an appalling state of affairs, no matter how critical one might be of conventional α values.

Table 4 does cast some light on the problem that ended the section on Fisher's LSD. It is true that, for our Table 2 data (which had N = 6), H_0 at {7} will be correctly rejected in about 54 of 100 experiments (see $\beta S = 0.5373$), but there are still 10 values of μ_j (j = 3 to 12) that are equal to one another.

Assuming that m_1 is the largest mean and comes from the large μ_1, and m_2 the smallest mean from the small μ_2, then m_1-m_2 will be the largest difference, and there will be 10 means remaining (being, $\mu_3 = \mu_4 = \ldots \mu_{12}$). Table 4 tells us that the probability of rejecting the true null (by a *t*-test) for a pair over that span of 10, is 0.602, and moving in to a span of 5 means, the probability of erroneous rejection for a true null over a 5-span pair will be 0.278. The fact that this kind of fiasco may happen in only 54 of 100 experiments should provide little comfort to the Fisher LSDer!

Table 4

True Type-1 Error Rate For $t_{0.05;\nu_2}$ (N = 6)

J	=	6	12	24
ν_2	=	30	60	120
$t_{0.05,\nu_2}$	=	2.042	2.000	1.980
Q	=	2.888	2.829	2.800
Range	=	6	12	24
True Type-1	=	0.344	0.691	0.949
Sub-Range	=	4	10	12
True Type-1	=	0.196	0.602	0.706
Sub-Range	=	3	5	6
True Type-1	=	0.119	0.278	0.359

Methods for Testing Planned Contrasts

There are three well known approaches to testing (J-1), or fewer, linearly independent contrasts that were planned independently of the sample data used to test them.

Properly Planned *t*-tests: This is the method usually referred to as 'the method of planned contrasts'. In that method, no more than H = J-1 linearly independent contrasts are chosen before the sample data are collected (or CERTAINLY without knowing anything about how the sample data turned out), then testing and deciding whether to accept or reject each of those null contrasts after the data are collected. The temptation to treat a contrast, chosen after looking at the data, as if it was a planned contrast, MUST be resisted because that greatly increases the type-1 error rate, as Table 4 shows. It will not do to say, "Well, I could easily have planned this contrast, now that I think about it!"

Properly done, planned *t*-tests can be very effective. For our Table 2 data, if a researcher chose to test $\mu_1 - \mu_2 = 0$, among possibly J-2 further contrasts, he or she would detect the falsity of that null with probability 0.8764 (if N = 6 is being used). Of course, tests on any of the μ_k (for k > 2) would be tests on true nulls and those J-2 should show null rejection at the rate $\alpha = 0.05$, if the contrasts had been chosen prior to the data collection. However, if our researcher did not 'know' the 'wise' tests (usually based on very good prior evidence - clairvoyance is not reliable!) and, from among the (say) J = 6 groups used, he or she tested $\mu_1 - \mu_3 = 0$ instead of $\mu_1 - \mu_2 = 0$, that smaller difference (70-60 = 10 rather than 70-50 = 20), would be detected with a probability of only 0.3886 (rather than the 0.9179 for $\mu_1 - \mu_2 = 0$ with this 6-groups larger $\nu_2 = 30$). There is always an infinite number of sets of H = J-1 linearly independent contrasts possible, among J > 2 means; so 'wise' choice is either very wise or very 'suspicious'. More likely, the honest choice misses some of the most seemingly interesting results, and finding one or two F_h fairly close to the critical $F\alpha;1,\nu_2$ but not large enough to reject the null may engender some mental stress - at the very least!

Planned Contrasts with Experimentwise Error Rates: There are at least two procedures for testing H = J-1 linearly independent, planned contrasts, but with control of the experimentwise type-1 error rate at a conventional level (e.g., 0.05, 0.01). These are the use of Bonferroni's Inequality, and Dunnett's (1964) Many-one procedure. It is appropriate here to ask how well these methods would detect the true differences in our illustration data in Table 2.

First, just as we might interpret Duncan's method for the *post hoc* evaluation of comparisons as an application of the multiplication theorem of probabilities, Bonferroni’s Inequality uses the addition theorem of probabilities. That says, in effect, if you wish to test H null contrasts and ensure that the probability of rejecting any one (or more) of them in error does nor exceed 0.05, you should not use the individual rate 0.05 but rather use 0.05/H. Thus when J = 6 and H = 5, the notional 0.05 is replaced by 0.01. Also, when J = 12, H = 11, then 0.05 is replaced by 0.05/11 = 0.004545; and so on. Those changes will reduce considerably the rate for detecting nulls that are not true. Dunnett's Many-one procedure is for H = J-1 comparisons, in which each of J-2 means is compared to one particular one, e.g., a control group. Dunnett (1964) produced special tables of his *t*-like statistic $d\alpha;H,\nu_2$ by integrating multivariate *t* distributions.

As with all planned contrasts, it matters greatly for the Bonferroni procedure which contrasts are planned. In our Table 2, $\mu_1 - \mu_2 \neq 0$ will be much more often detected than μ_1-μ_3, while null contrasts across the μ_k (k > 2) are all true; so will be rejected rarely - and ever more rarely as J increases. A similar principle applies to Dunnett's Many-one method. To find out how the μ_j might be patterned, it matters greatly which 'control' mean is selected. Obviously μ_1 would be the most revealing because then we would be testing $\mu_1 - \mu_2$, $\mu_1 - \mu_3$, $\mu_1 - \mu_4$, $\mu_1 - \mu_5$, $\mu_1 - \mu_6$ and so on. That should show a big difference for the first, and a smaller difference for each of the others. But if the 'control' mean is any of the μ_k for k > 2, then there will be J-3 null differences, and two moderate differences (with μ_1 and with μ_2).

The use of Bonferroni-adjusted *t*-tests has been advocated, and seemingly is still in use, for contrasts chosen *post hoc*. But Rodger (1973) showed that *post hoc* choice inflates the experimentwise type 1 error rate beyond the claimed α.

The results of the numeric integrations of the noncentral variance ratio distribution are shown in Table 5. The critical *t* values used have been squared to make F values - FB for Bonferroni's *t*, FN for Dunnett's *t*. Also, the results of a conventional 5% *t*-test are given for Ft. The power (β) for rejecting two or three particular null contrasts (assuming they are among the J-1 planned) are given, βB for Bonferroni's method and βN for Dunnett's procedure, and βt for Ft. Only two comparisons are shown for Dunnett's (comparison-restricted) method. The μ_j for the planned contrasts are those given in Table 2, and these make Δ_h = 12.0, 9.0, and 3.0 for the three stated contrasts, respectively.

As is well known, Dunnett's method has a bit more power than Bonferroni's procedure. However, Dunnett's method is restricted to a particular set of comparisons (the many-one set). Also the power for detecting that the largest comparison μ_1-μ_2 is not zero, though apparently better than either the Scheffé or Tukey (NKMR) methods, is notably poorer than what either Duncan's or Rodger's method could produce. And that does not seem to be well known. Also, Rodger's procedure gives one a free, *post hoc*, choice of contrasts (subject only to the r rule of {27} and the requirement of linear independence, preferably mutual orthogonality).

The detection power of both Bonferroni's method and Dunnett's depends on how 'wisely' the planned contrasts (or 'control group') have been chosen - this is not a problem for Duncan, and especially not for Rodger.

Table 5

β for Specified, Planned Contrasts

Method		J = 6	12	20
For N = 6,	ν_2 =	30	60	100
Bonferroni	FB =	7.563	8.694	9.518
$(\mu_1-\mu_2)$	βB =	0.7565	0.6950	0.6473
$(\mu_1+\mu_3-2\mu_2)$	βB =	0.6018	0.5246	0.4699
$(\mu_1-\mu_3)$	βB =	0.1741	0.1224	0.0943
Dunnett	FN =	7.398	8.180	8.820
$(\mu_1-\mu_2)$	βN =	0.7655	0.7245	0.6883
$(\mu_1-\mu_3)$	βN =	0.1811	0.1401	0.1143
Conventional	Ft =	4.171	4.001	3.936
$(\mu_1-\mu_2)$	βt =	0.9179	0.9262	0.9293
$(\mu_1+\mu_3-2\mu_2)$	βt =	0.8271	0.8393	0.8440
$(\mu_1-\mu_3)$	βt =	0.3886	0.3993	0.4035

Note: $\mu_1 = 70$, $\mu_2 = 50$, $\mu_3 = \mu_4 = \ldots = \mu_J = 60$, σ^2=100, N=6 as for Table 2.

βt improves as one moves across the columns, in the Ft section of Table 5. That is because increasing ν_2 increases power. But that phenomenon is swamped for FB and FN because increasing J increases H = J-1, decreases the decision error rate (to keep the experimentwise error rate constant at a conventional value) and thereby reduces the contrast power β. Also, for each of the procedures, as one moves down a column, power diminishes. That is because Δ_h diminishes from 12 through 9 to 3.

Finally, when J = 6, according to Table 2, Duncan's DMR had $\beta D = 0.9144$, and Rodger's $\beta R = 0.9059$. Both of these are notably better than either Bonferroni's or Dunnett's method, and almost as good as the planned *t* for $\mu_1 - \mu_2 = 0$. Duncan's βD is for m_{LG}-m_{SM}, the largest observed difference, very likely for $\mu_1 - \mu_2$. Rodger's βR uses $6\Sigma(m_j\text{-}m.)^2$, which is an estimate of $6\Sigma(\mu_j - \mu.)^2$, and for our data in Table 2, is $= 6\Sigma(\mu_j - \mu.)^2 = 6(100+100) \equiv 6(70\text{-}50)^2/2 = 6(\mu_1\text{-}\mu_2)^2/\Sigma c^2_j$. With both of these *post hoc* methods, one has the huge advantage of being informed about how the m_j turned out and, with Rodger's procedure, complete freedom to choose linearly independent contrasts of any form (not just comparisons).

Unfortunately, rather few statistical investigations of contrasts set g and β or $E\beta$ before the data are collected, and most fail to compute the sample size N required to make the g detection rate β or $E\beta$. For these reasons science progresses more slowly than it should. It is hoped that this paper will make such preparation easier and less confusing. If N were chosen to detect a stated g at a reasonable rate β or $E\beta$, there would be fewer ups and downs among research reports. What the investigator claims to have found and what not to have found would be clearer; so easier to be challenged or refuted by others. Surely, that is the way for scientific knowledge to progress.

Concluding Statement

This report has examined the 'power' (β) various methods have to detect false null hypotheses, such as the traditional (overall) H_o, and null contrasts, in fixed-effects statistical investigations of the 'true' means (μ_j) of normal variates of J populations. The methods studied were either designed to evaluate contrasts ***post hoc*** or as a **pre-planned** set. The methods also either employ a conventional type 1 error rate (e.g., $\alpha = 0.05$) on an **experimentwise basis** (for the J populations), or on a **decision basis** (e.g., expected rejection rate $E\alpha = 0.05$) for the final decisions about H = J-1 contrasts. The extent of the falsity of H_o was fixed by using a constant noncentrality parameter (Δ_m). This came from using a constant sample size (N) with a constant pattern and amount of variation among the true means (μ_j). In that way there was no confounding of effect size (Δ_m) and investigation size (J populations). The results are summarized below.

I: For *post hoc* methods with conventional, experimentwise error rates (Scheffé, Tukey and Newman-Keuls), the power (β) for detecting the falsity of H_0 drops dramatically as J increases.

II: That same type of detection loss for those three procedures occurs in their ability to detect false null contrasts.

III: Those losses of power (and of false null contrast detection capacity) with increasing J, do not occur for *post hoc* methods that use decision-based error rates (Duncan and Rodger).

IV: The above findings are true whether the null test procedure uses the variance ratio distribution or the Studentized range distribution.

V: Though not, strictly speaking, a power matter, discrimination among the true μ_j is hampered if null contrasts are never accepted. Of course, null acceptance has to be at a reasonable degree of approximation, and that requires computing (then using) the sample size (N) necessary to detect a certain amount of null falsity (say, |g|, as described in this paper), with fairly good probability (e.g., Eβ = 0.95).

VI: Limiting *post hoc* contrast testing to comparisons only (as required in the multiple range methods of Duncan and of Newman-Keuls) either reduces true discrimination between the means (μ_j) or requires considerable increases in sample size (N) to get around that problem.

VII: Testing H = J-1 pre-planned, linearly independent contrasts (it is mathematically impossible to have more than J-1 contrasts that are linearly independent of one another) with a conventional, decision-based type 1 error rate (especially against specific-sized alternatives), works quite well if the contrasts and their alternatives were chosen 'wisely'. That is, the false nulls each have a reasonably-sized value (say, g^2), and sample size N was computed to give fairly good probabilities of false null detection.

VIII: The temptation to "vary one's pre-planned contrast choices" after the test data have been examined can be very high, in the light of seeing some 'now obvious' choices that one "could easily have planned"! To succumb to that temptation amounts to indulgence in *post hoc*, unplanned *t*-tests. That yields very large type 1 error rates (beyond the conventional α asserted), and those unplanned error rates grow ever larger as J increases. The results of unplanned, *post hoc* *t*-tests continue to be published with seemingly little concern for the large 'actual' type 1 error rates, that are demonstrated in this paper.

IX: Testing H = J-1 pre-planned contrasts with a conventional experimentwise type 1 error rate (e.g., $\alpha = 0.05$), as with the Bonferroni or the Dunnett Many-one procedure, reduces the probability of detecting a false null contrast considerably (compared to simple, conventional *t*-tests) - though not as badly as the *post hoc* methods of Scheffé or Tukey. And that loss of detection probability grows worse as J increases. But even those facts depend on having chosen one's pre-planned contrasts 'wisely', i.e., choosing those contrasts that have sufficient values of g^2 to yield respectable probabilities of detection.

X: The last method considered was Fisher's LSD procedure. This *post hoc* method uses a mixture of conventional, experimentwise type 1 error rate (e.g., $\alpha = 0.05$) and a 'supposed' conventional decision-based type 1 error rate (e.g., supposedly $\alpha = 0.05$). The use of the experimentwise basis, exactly like the Scheffé method, makes the probability of detecting a false null contrast diminish as J increases. If the method rejects H_0 because $F_m \geq F\alpha;\nu_1,\nu_2$ then *t*-tests are used to choose which contrasts to declare 'significant' - using, in effect, $t^2 = F\alpha;1,\nu_2$. That amounts to unplanned *t*-tests on the means; so true null contrasts across the μ_j (that still remain) are at high risk of erroneous null rejection (well beyond the 'supposed' α asserted). The pattern of reduced detection rates for some false null contrasts, combined with greatly elevated type 1 error rate for other true null contrasts, is not a type of yo-yo procedure that recommends itself!

XI: Planned *t*-tests, Duncan's method and Rodger's procedure are the only forms of analysis that unpick true differences among the μ_j (using the sample m_j) at a reasonable rate, for moderate-sized effects, with practical sample sizes (N). But Duncan's method is restricted to comparisons, and that severely limits the true differences it can find. Rodger's scale-free, noncentrality parameter g (see {2}, {6}, {15}, {28} and Table 3) makes the setting of power (or non-null detection rate) easy. One can pre-set g for alternatives to null contrasts, even before one knows what contrasts will be decided *post hoc*. Making the choice of g and Eβ when designing a study, and computing sample size N (see {28}) to yield that Eβ is a procedure that is strongly recommended. Of almost equal importance is the recommendation that the pre-chosen g and Eβ be reported along with the results of one's research. In that way, readers will not only learn what was found, but also how much was sought, at what rate, and what was not found. Those things would be a great help in interpreting reported findings, in designing follow-up studies, and in reducing ineffective further studies.

Author Notes: Dr. Rodger is a post-retirement professor of psychology at Dalhousie University, Halifax, Nova Scotia and may be contacted at rodgermethod@gmail.com

References

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8,* 1–62.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences (2nd ed.).* Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 135-159.

Duncan, D. B. (1951). A significance test for differences between ranked treatments in the analysis of variance. *The Virginia Journal of Science, 2* (n.s.), 171-189.

Duncan, D. B. (1952). On the properties of the multiple comparisons test. *The Virginia Journal of Science, 3* (n.s.), 49-67.

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

Dunnett, C. W. (1964). New tables for multiple comparisons with a control. *Biometrics, 20*, 482-491.

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver & Boyd.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*, 107-128.

Keuls, M. (1952). The use of studentized range in connection with an analysis of variance. *Euphytica, 1*, 112-122.

McDowell, M. A., Fryar, C. D., Ogden, C. L., Flegal, K. M. (2008). *Anthropometric Reference Data for Children and Adults: United States, 2003-2006*. National Health Statistics Reports (10).

Morrison, S. E. (2004). The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychological Methods*, *9*, 147-163.

Newman, D. (1939). The distribution of range in samples from a normal population, expressed in terms of an independent estimate of standard deviation. *Biometrika, 31,* 20-30.

Neyman, J., Pearson, E. S. (1928a). On the use and interpretation of certain test criteria for the purpose of statistical inference: part I. *Biometrika, 20A*, 175-240.

Neyman, J., Pearson, E. S. (1928b). On the use and interpretation of certain test criteria for the purpose of statistical inference: part II. *Biometrika, 20A*, 263-294.

Neyman, J., Pearson, E. S. (1933a). On the testing of statistical hypotheses in relation to probabilities *a priori*. *Proceedings of the Cambridge Philosophical Society, 29*, 492-510.

Neyman, J., Pearson, E. S. (1933b). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society, London,* 29, 289-337.

Roberts, M. (2011). Simple, powerful statistics: an instantiation of a better 'mousetrap'. *Journal of Methods & Measurement in the Social Sciences, 2*, 63-79.

Rodger, R. S. (1967a). Type I errors and their decision basis. *British Journal of Mathematical and Statistical Psychology, 20*, 51-62.

Rodger, R. S. (1967b). Type II errors and their decision basis. *British Journal of Mathematical and Statistical Psychology, 20*, 187-204.

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

Rodger, R. S. (1973). Confidence intervals for multiple comparisons and the misuse of the Bonferroni inequality. *British Journal of Mathematical and Statistical Psychology, 26*, 58-60.

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. (1976). Tables of Stein's non-central parameter D, required to set power for numerical alternatives to H_0 tested by two-stage sampling anova. *The Journal of Statistical Computation and Simulation, 5*, 1-22.

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.

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

Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics, 16*, 243-258.

Tukey, J. W. (1953). "The Problem of Multiple Comparisons." Privately circulated, dittoed manuscript.

[A] Consider having K (e.g., K=12) samples, each from a separate sub-population or each having been given a different 'treatment', and (for convenience) arrange the means in order of size, m_1 the smallest to m_K the largest. If we test the null hypothesis that all the true means μ_k are equal to one another by analysis of variance, rejecting that null if the overall, observed $F_m \geq F\alpha;K\text{-}1,\nu_2$, then α is the '**experimentwise type 1 error rate**', and α is conventional if it is, e.g., 0.05 or 0.01. If the K samples belong to I×J (e.g., 4×3) sub-classes, a factorial analysis is then the popular form of analysis. That evaluates the I main effect, the J main effect, and the IJ interaction, rejecting the respective nulls if the observed $F_I \geq F\alpha;I\text{-}1,\nu_2$, $F_J \geq F\alpha;J\text{-}1,\nu_2$, $F_{IJ} \geq F\alpha;(I\text{-}1)(J\text{-}1),\nu_2$. The α used here is the '**familywise type 1 error rate**' for the I, the J and the IJ families, respectively. Also, when the Newman-Keuls multiple range method **'steps in'** to test $\mu_{11}\text{-}\mu_1 = 0$ and $\mu_{12}\text{-}\mu_2 = 0$ against $q\alpha;11,\nu_2$, that α is the '**familywise type 1 error rate**' for the 11-groups families, and it is conventional because Newman-Keuls uses, e.g., α = 0.05 or 0.01. For Duncan the test criterion for the two sub-range comparisons would be the unconventional $q\gamma;11,\nu_2$ with $\gamma = 1\text{-}0.95^{10} = 0.40$. The formula $\gamma = 1\text{-}0.95^{10}$ is a standard that is based on the probability (0.05) of making a type 1 error in 10 statistically independent test decisions (though Duncan's comparisons are not generally statistically independent of one another); so Duncan's method actually uses a '**decision-based, familywise type 1 error rate**' for the 11-groups families. Suppose the investigator had a very, very good idea about possible values of the μ_k and was therefore able to plan to test a wise set of K-1 =11 linearly independent null contrasts across the μ_k, each to be tested by a two-tailed *t*-test (or its equivalent t^2 = $F\alpha;1,\nu_2$). That investigator is using a '**decision-based type 1 error rate**'. The probability of a type 1 error will be α for each of the 11 *t*-test decisions (for true nulls). Finally, the Rodgerian would compute r = $[F_m/F[E\alpha];11,\nu_2] \leq 11$, then look through the data to find r rejectable null contrasts that each satisfy $F_h \geq F[E\alpha];11,\nu_2$. Those r nulls would be rejected and 11-r others (that did not reach the $F[E\alpha];11,\nu_2$ criterion) retained. All ν_1 = 11 contrasts in the decision set would be linearly independent of one another (preferably mutually orthogonal) and make reasonable scientific sense. That Rodgerian is using a '**decision-based type 1 error rate**' because, over a long series of such investigations in which all the μ_k are truly equal, the average of this investigator's rate of type 1 error (i.e., the average ratio r/ν1) will be Eα. Experimentwise and familywise error rate (α) is the area in the tail of a distribution (or in two tails for two-tailed tests). Rodgerian decision-based error rate (Eα) is the weighted average of successive probabilities (of r = 0, 1, . . . , ν_1) in the F distribution (see Rodger, 1975a, p. 76 Figure 2 for a diagram and a numeric illustration; http://en.wikiversity.org/wiki/Rodger's_Method for more numbers; and Rodger, 1975b, p. 230 Figure 1 for diagrams on both Eα and Eβ). Taking a distribution by its tail (α) is a procedure that can be somewhat unstable (non-robust) when all the assumptions are not quite met, but grasping the distribution around its middle (for Rodger's Eα and Eβ) is more stable – what's not to love about that?

[B] The language used here (and for most other outcomes described in this paper) to say how investigators report on their statistical evaluation of null hypotheses (or null contrasts) is very general (and therefore somewhat vague). That is because there is wide variation between how investigators make those reports, ranging from statistical decisions that say what the investigator believes her/his data indicates is true (at least to some, reasonable degree of approximation), all the way to reporting the data themselves with little analytic interpretation. Of course, we all know that statistical analysis is subject to error, but that should not preclude the investigators (who are most familiar with their data, and how it was collected) from saying what they believe their data demonstrate. Surely that is more likely to encourage scientific progress.

[C] The fact that using conventional, experimentwise error rates (e.g., α = 0.05 or 0.01) results in a notable loss of power (β), for fixed Δ_m, as J (the number of 'treatment' groups) is increased, is a feature of 'fixed effects' statistical designs and analysis. Exactly the opposite happens with 'random effects' (or 'variance components') statistical designs and analysis. Those show a notable increase in power (β), for fixed N and 'treatment' variance σ^2_τ, as J is increased. Power for 'fixed effects' is an integral of the noncentral F distribution. Power for 'random effects' is an integral of the central F distribution. For example, if we have J = 4 'treatment' temperatures (say, 5°C; 12°C, 14°C and 27°C), a 'fixed effects' design has chosen those values deliberately, and the analyst wants to know which of them differ from which in their measured effect on the variate, and by how much. In a 'random effects' design those particular, four temperatures were drawn at random from, say, 0°C to 30°C. That analyst has no particular interest in the four treatment values that randomization popped up. Her/his interest is in the amount of variation (σ^2_τ), if any, in the variate measurements that temperature differences generate. Notice that the words "if any" constitute a reasonable question when you are unsure whether the 'treatment variable' has any effect worth mentioning; so experimentwise error rate is a reasonable criterion. One wonders whether the popular use of (ill-advised) experimentwise error rates (in 'fixed effects' analyses) is a holdover from historical confusion!

[D] Regarding 'how small is small'; suppose we wanted to compare the average standing heights of two sub-populations. One would be a particular type of adult human males and the other that same type of adult human females. We will use a conventional *t*-test, with $\alpha = 0.05$, to decide on the $\mu_m - \mu_f$ comparison, and we wish to have the probability $\beta \geq 0.95$ of detecting if the two sub-population means (the μ's) differ by at least $|0.1|$ inches (i.e., about 0.25cm). The data from the McDowell et al. (2008) study, used in Table 3, suggests that $\sigma \approx 11$ cm. Since $\Delta 0.95;1,\nu_2 \approx 13$ for large ν_2, we would require each of our two, independent, random samples to be of size $N \approx 50{,}000$. The small difference sought ($|g| = 0.016$) would have to be quite important, scientifically, to warrant so much work and money. Undoubtedly, there are occasions when $\alpha = 1\text{-}\beta$ is not appropriate, but such circumstances are very much a subject-matter concern, not (in general) something statistics can settle in principle.

[E] Cohen (1988, 1992) had been calling upon behavioural scientists since the early 1960s to address the matter of power in their research designs, but without much success. In his book (1988) he dealt with at least eight statistical procedures and had a 'different' measure of noncentrality for each of them. He also provided separate 'power tables' for each of these. The richness of his explanations may have be rather confusing for the relatively casual, occasional scientific user. It is therefore unfortunate on that account, but the opportunity was lost to show that all these statistical methods use the same noncentral F distribution, with a standard noncentrality parameter. For example, Cohen's parameter d (for *t*-tests on mean differences) is related to Rodger's, scale-free g given here by $g^2 = d^2/2$. He defined d = 0.2, 0.5, 0.8 to be 'small', 'medium' and 'large' effect sizes. In g terms those are g = 0.14, 0.35, 0.57 respectively. For controlled experimental studies, these are all rather small effects: even Cohen's 'large' effect size is not very big. But Cohen's definitions might be more reasonable for population surveys.