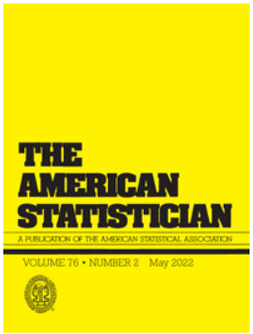
Taylor & Francis
Taylor & Francis Group



The American Statistician

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

Comparing Three Groups

Jelle J. Goeman & Aldo Solari

To cite this article: Jelle J. Goeman & Aldo Solari (2022) Comparing Three Groups, The American Statistician, 76:2, 168-176, DOI: 10.1080/00031305.2021.2002188

To link to this article: https://doi.org/10.1080/00031305.2021.2002188

THE AMERICAN STATISTICIAN
2022, VOL. 76, NO. 2, 168–176: Statistical Practice
https://doi.org/10.1080/00031305.2021.2002188

Comparing Three Groups

Jelle J. Goeman[a] and Aldo Solari[b]

[a]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands; [b]Department of Economics, Management and Statistics, University of Milano-Bicocca, Milan, Italy

ABSTRACT

For multiple comparisons in analysis of variance, the practitioners' handbooks generally advocate standard methods such as Bonferroni, or an *F*-test followed by Tukey's honest significant difference method. These methods are known to be suboptimal compared to closed testing procedures, but improved methods can be complex in the general multigroup set-up. In this note, we argue that the case of three-groups is special: with three groups, closed testing procedures are powerful and easy to use. We describe four different closed testing procedures specifically for the three-group set-up. The choice of method should be determined by assessing which of the comparisons are considered primary and which are secondary, as dictated by subject-matter considerations. We describe how all four methods can be used with any standard software.

ARTICLE HISTORY
Received July 2020
Accepted October 2021

KEYWORDS
Analysis of variance; Closed testing; Dunnett; Multiple comparisons; Multiple testing; Tukey's honest significant difference

1. Introduction

Researchers often compare an outcome measure between several experimental or observational groups. Designs with two groups are most common, and are well discussed in statistical handbooks. Multi-group designs are more complex, since they give rise to many between-group comparisons. If more than one such comparison is of interest, then multiple testing problems arise, and adjustments need to be made to prevent excessive false positive results.

Practical statistical handbooks often touch upon multiple comparisons methods when discussing post hoc testing in analysis of variance (ANOVA) to determine which of the groups are different after a significant ANOVA test. However, they tend to focus on methods that are generally applicable in the multigroup case. The most frequently advocated methods are Tukey's honest significant difference (HSD), or Bonferroni, usually after the ANOVA test (e.g., Field 2013; Glover and Mitchell 2008; Stevens 2013; Tabachnick, Fidell, and Ullman 2007).

In the specialized literature, it is known that these methods are suboptimal, in the sense that they may be uniformly improved by methods based on closed testing (Marcus, Peritz, and Gabriel 1976; Goeman, Hemerik, and Solari 2021). Closed testing methods always reject at least as many hypotheses, and possibly more, while still controlling the same error rates. However, these methods can be computationally and conceptually complicated in the general multigroup setting, and are not usually implemented in standard software packages (Begun and Gabriel 1981; Bergmann and Hommel 1988; Rom and Holland 1995).

In this article, we argue that the case of three groups, arguably the most common multigroup design, deserves special treatment. In the three-group case, the improved closed-testing-based methods are relatively easy, and within reach of standard software packages. The power improvements of such methods over, for example, ANOVA followed by Tukey's HSD can be substantial. We present four different closed testing-based approaches for the three group design. We will argue that the choice for one of these methods should be based on an a priori choice which of the comparisons are of primary and which are of secondary interest. We will initially concentrate on the well-studied one-way ANOVA design with three equal size groups, but extend to other parametric and nonparametric setups in Section 10. Example analyses are provided in a simple and a more complex model in Section 11 and 12. We briefly touch upon extension of these ideas to the four-group situation in Supplementary Material A.

The subject of multiple comparisons has a long history and a huge literature, which we cannot cover in full (see e.g., Miller 1981; Hochberg and Tamhane 1987; Hsu 1996; Bretz, Hothorn, and Westfall 2011; Dickhaus 2014; Cui et al. 2021). This article aims to give practical guidelines for users in a single particular but important situation. However, the discussion of this special case touches upon many of the central issues in multiple testing, and we hope that our note may serve as a gentle introduction to the wider subject.

2. Four Hypotheses for Three Groups

We assume a three-group design with a parameter of interest per group, denoted μ_1, μ_2, and μ_3. These parameters may be

CONTACT Jelle J. Goeman j.j.goeman@lumc.nl Leiden University Medical Center, Leiden 2300 RC, Netherlands.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/TAS.

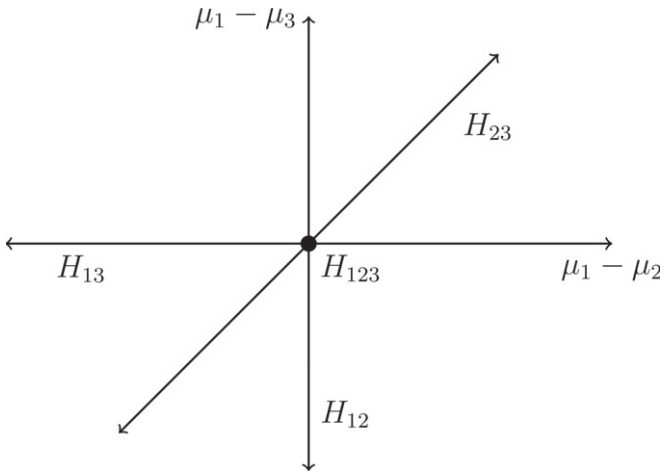


Figure 1. Visualization of the four hypotheses H_{12}, H_{13}, H_{23}, and H_{123} in a parameter space with axes $\mu_1 - \mu_2$ and $\mu_1 - \mu_3$. Note that H_{23} is the diagonal line for which $\mu_1 - \mu_2 = \mu_1 - \mu_3$. In the origin all four hypotheses are true; elsewhere at most one.

means (in the classical ANOVA setting), proportions (in a 2×3 contingency table), per-group regression coefficients, or even distributions (in nonparametric settings). To keep the discussion concrete, we will focus on the one-way ANOVA setting with equal-size groups as the central example, coming back to the general case in Section 10. In this classical setting, we assume we have three groups of n observations. Each observation is normally distributed around its group mean μ_1, μ_2, or μ_3, with common variance σ^2.

We may formulate four null hypotheses to compare the group means μ_1, μ_2, and μ_3. First, the so-called global null hypotheses that all three-group means are equal

$$H_{123}\colon \mu_1 = \mu_2 = \mu_3.$$

Next, there are the three pairwise comparisons between groups

$$H_{12}\colon \mu_1 = \mu_2; \qquad H_{13}\colon \mu_1 = \mu_3; \qquad H_{23}\colon \mu_2 = \mu_3.$$

The four hypotheses H_{123}, H_{12}, H_{13}, and H_{23} are logically related to each other: if any two are true, then all must be true. For example, if H_{12} and H_{13} are true, then $\mu_1 = \mu_2$ and $\mu_1 = \mu_3$, so that we have $\mu_1 = \mu_2 = \mu_3$, which implies that H_{123} and H_{23} are also true. The number of true hypotheses among H_{123}, H_{12}, H_{13}, and H_{23} can therefore be either 0, 1, or 4, but never 2 or 3. Additionally, if only one hypothesis is true, this cannot be H_{123}. These logical implications between hypotheses are also known as restricted combinations (Shaffer 1986), and are visualized in Figure 1.

3. The Need for Multiple Testing Correction

A false-positive result, a rejection of a true null hypothesis, may result in an incorrect scientific finding reaching the scientific literature, and should therefore be prevented. The convention is to accept a probability of such a false positive result of at most $\alpha = 0.05$, which is achieved by in single hypothesis testing by only rejecting hypotheses with a p-value below α.

If multiple hypotheses are tested, then each hypothesis again has a probability α of a false positive result. Therefore, without adjustment the probability that at least one a false-positive result occurs as a result of the experiment tends to exceed the acceptable rate. For example, in the ANOVA set-up with three equal size groups of 10 at $\alpha = 0.05$, simply testing all four hypotheses at $\alpha = 0.05$ results in an excessive 13% of experiments producing at least one false positive result in the situation that all hypotheses are true (see Supplementary Material B).

We can correct for multiple comparisons by controlling the *familywise error rate* (FWER), the probability of obtaining one or more false positive results (Tukey 1953). Methods controlling FWER bring the probability of producing at least one false positive result back to at most the required α level. Alternatively stated, they guarantee that at least $(1 - \alpha) \times 100\%$ of three-group experiments performed produce no false-positive results. We always consider control of FWER in the strong sense, that is, control must hold for all possible values of μ_1, μ_2, and μ_3. A FWER guarantee is necessary if results are selectively emphasized, for example, in the discussion, title or abstract of articles. Researchers may unknowingly emphasize the—surprising—false-positive results. This will result in excessive false positive rates among the emphasized results if FWER control was not applied (Benjamini 2019).

4. Primary and Secondary Hypotheses: Four Scenarios

The four hypotheses given above are seldom of equal interest to the researcher, and we may distinguish between hypotheses of primary and secondary interest. Hypotheses of primary interest are those that are central to the research question; hypotheses of secondary interest are those that become of interest only as a follow-up to the primary research question (Bretz et al. 2009; Burman, Sonesson, and Guilbaud 2009). The decision which hypotheses are primary and which are secondary should always be based on subject-matter knowledge, independent of the data.

For the three-group comparison case, we distinguish four scenarios for the choice of primary and secondary hypotheses.

(A) *The global hypothesis H_{123} is primary*: This is natural when the presence of any difference between the means can directly be meaningfully interpreted, regardless of the location of such difference. For example, consider the case that the three groups represent three levels of an ordinal variable, created by categorizing a numerical outcome. In this case, if H_{123} is false, we can say that the ordinal variable (and consequently the underlying numerical one) is associated with the outcome, even if we don't yet know what the form the association takes. Similarly, in a genetic study where the three groups are three phenotypes *AA*, *Aa* and *aa*, rejection of H_{123} is sufficient to establish the presence of a genetic effect. The pairwise hypotheses are secondary, giving additional information on the mode of inheritance.

(B) *All three pairwise hypotheses, H_{12}, H_{13}, and H_{23}, are primary*: This is natural when the three groups represent categories of a nominal variable, and all three groups are equally important. In this case, it is usually not satisfactory just to know that some group differences exist, but researchers would always want to know specifically *which*

groups differ. For example, if subjects would be of three different nationalities, just knowing that there is an association between nationality and outcome is not very informative; the researcher would always want to know which nationalities differ from each other.

(C) *Two of the pairwise hypotheses, say H_{12} and H_{13}, are primary*: This is natural when Group 1 represents a reference against which both other groups are compared. In this case, the comparison of Groups 2 and 3 only becomes of interest when at least one of the groups has been shown to be different from the reference. A classical example of this situation is the case when two novel treatments are compared against a placebo: the researcher is only interested in a difference between the treatments if at least one has been shown to outperform the placebo.

(D) *One of the pairwise hypotheses, say H_{12}, is primary*: This is natural when one of the groups (Group 3) is of secondary interest. For example, Groups 1 and 2 could be placebo and high dose treatment, while Group 3 is a low dose treatment. In this case the researcher could be only interested in the relative effect of the low dose after the effectiveness of the high dose has been established.

5. Standard Methods

The standard tests for H_{123}, H_{12}, H_{13}, and H_{23} in the ANOVA model are the (partial) F-tests. The information from the observations on these hypotheses is summarized in the estimates $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\mu}_3$, and pooled variance estimate $\hat{\sigma}^2$. For the discussion in this article, it is useful to remark that the partial F-tests can be rewritten to equivalent tests based on the standardized group differences. Ignoring multiplicative constants, the partial F-test statistic for H_{12} is proportional to the standardized squared group difference

$$S_{12} = \frac{(\hat{\mu}_2 - \hat{\mu}_1)^2}{\hat{\sigma}^2};$$

analogous for H_{13} and H_{23}. The distributions of S_{12}, S_{13}, S_{23} are identical under the null hypotheses; let c_α be the $1-\alpha$-quantile of that distribution. For H_{123} the F test is proportional to the test statistic

$$S_{123} = S_{12} + S_{13} + S_{23}, \tag{1}$$

as shown in the supplemental material A. Let c_α^{123} be the $1-\alpha$-quantile of the distribution of S_{123}.

To control FWER, the most frequently recommended solutions are Bonferroni and Tukey's honest significant difference (HSD; Tukey 1949). With Bonferroni, instead of rejecting each H_{ij} when $S_{ij} \geq c_\alpha$, we reject when $S_{ij} \geq c_{\alpha/3}$, adjusting the α-level by a factor 3 to adjust for the three comparisons. Tukey's HSD method rejects when $S_{ij} \geq \tilde{c}_\alpha$ instead, where $\tilde{c}_\alpha$ is the $(1-\alpha)$-quantile of the distribution of

$$\tilde{S}_{123} = \max(S_{12}, S_{13}, S_{23}),$$

which is proportional to a studentized range distribution (Tippett 1925). Tukey's HSD method is uniformly more powerful than Bonferroni, since Bonferroni's quantile $c_{\alpha/3}$ is an upper bound to the quantile of $\max(S_{12}, S_{13}, S_{23})$ among all possible joint distributions of three statistics S_{12}, S_{13}, and S_{23}, while $\tilde{c}_\alpha$ is the same the quantile calculated using the specific joint distribution of the statistics in the ANOVA model. We have $c_{\alpha/3} > \tilde{c}_\alpha > c_\alpha$.

Tukey's HSD or Bonferroni may be applied directly on H_{12}, H_{13}, and H_{23}, without first looking at the ANOVA F-test. However, these methods are often used as *post hoc* tests, to be performed only after the global ANOVA F-test rejects. The resulting two-step procedure, which we refer to as Tukey's post hoc procedure (analogous for Bonferroni) is as follows:

1. If $S_{123} \geq c_\alpha^{123}$, reject H_{123}.
2. If H_{123} was not rejected, stop; otherwise, reject each of H_{12}, H_{13}, and H_{23} for which the corresponding $S_{ij} \geq \tilde{c}_\alpha$.

Since it may happen that $S_{123} < c_\alpha^{123}$, while $\tilde{S}_{123} \geq \tilde{c}_\alpha$, Tukey's post hoc method has strictly less power than Tukey's HSD for rejecting each of H_{12}, H_{13}, and H_{23}. Conversely, however, it may also happen that $S_{123} \geq c_\alpha^{123}$, while $\tilde{S} < \tilde{c}_\alpha$; in that case the Tukey's post hoc procedure rejects H_{123} (only), while Tukey's HSD procedure would reject no hypotheses.

Most statistical software packages also offer Dunnett's procedure (Dunnett 1955), which is used analogously to Tukey's HSD procedure, but in the situation that only H_{12} and H_{13} are of interest. Dunnett's procedure rejects H_{12} and/or H_{13} when the corresponding test statistics exceed $\tilde{c}_\alpha^1$, where $\tilde{c}_\alpha^1$ is the $(1-\alpha)$-quantile of the distribution of

$$\tilde{S}_1 = \max(S_{12}, S_{13}).$$

Note that Dunnett's critical value is less stringent than Tukey's one, that is, $c_\alpha < \tilde{c}_\alpha^1 < \tilde{c}_\alpha$. While Dunnett's procedure controls FWER on H_{12} and H_{13} directly, is often used as a post hoc procedure, and embedded in a two-step approach after an ANOVA F-test, like with Tukey. Like Tukey's HSD, Dunnett's procedure is uniformly more powerful than Bonferroni's procedure on H_{12} and H_{13}, which would reject when S_{12} or S_{13} exceeds $c_{\alpha/2}$, since $c_{\alpha/2} > \tilde{c}_\alpha^1$.

6. Four Closed Testing Procedures

The multiple comparisons procedures described in the previous paragraph are not optimal, but can be uniformly improved by embedding them into a closed testing procedure (Marcus, Peritz, and Gabriel 1976). It is known (Sonnemann 2008; Goeman, Hemerik, and Solari 2021): that all FWER controlling procedures are either equivalent to a closed testing procedure, or can be uniformly improved by one. Closed testing procedures for the four hypotheses H_{123}, H_{12}, H_{13}, and H_{23} can be constructed as follows.

First, we must verify that the collection of hypotheses is closed with respect to intersection. That is, for every two hypotheses in the family we must ensure that their intersection is also in the family. An intersection between two hypotheses is a hypothesis that is true if and only if both intersected hypotheses are true. For example, the intersection between H_{12} and H_{13} is the hypothesis that $\mu_1 = \mu_2$ and $\mu_1 = \mu_3$, which is H_{123}, which is indeed in the family. It is easily verified using Figure 1 that the family of hypotheses H_{123}, H_{12}, H_{13}, and H_{23} is closed with respect to intersection.

In a closed family of hypotheses, the closed testing procedure controls FWER by requiring that any hypothesis is only rejected after all hypotheses implying it are rejected (Marcus, Peritz, and Gabriel 1976). One hypothesis implies another if the second hypothesis is always true when the first one is. For example, H_{123} implies H_{12} since $\mu_1 = \mu_2 = \mu_3$ implies that $\mu_1 = \mu_2$. We can see this implication in Figure 1, since H_{123} is a subset of H_{12}. The requirement that implying hypotheses must be rejected first is the only requirement closed testing imposes to achieve FWER control; all tests are performed at level α, without further α-level adjustment.

We see from Figure 1 that in the family H_{123}, H_{12}, H_{13}, and H_{23}, the hypothesis H_{123} implies all other hypotheses, but for the rest there are no implications between hypotheses. A closed testing procedure for these four hypotheses is therefore always a procedure in two steps. It first tests the implying hypothesis H_{123} at level α. If that hypothesis is not rejected, the procedure stops. Otherwise, in Step 2, all three of H_{12}, H_{13}, and H_{23} are tested, each at level α. All closed testing procedures in the three-group design, therefore, take this general form:

1. Test H_{123} with a valid α-level test
2. If H_{123} was not rejected, stop; otherwise, test each of H_{12}, H_{13}, and H_{23} with a valid α-level test.

At the first sight, the closed testing framework may seem quite restrictive. In fact, it is a very general framework, from which a great variety of methods can be constructed by choosing different options for a hypothesis test for each of the hypotheses. For FWER control, any valid α-level test may be chosen for any of the four hypotheses, as long as these tests are chosen independently of the data. The resulting procedures may have quite different power properties, depending on the chosen tests.

For the one-way ANOVA with three groups, we will construct four different closed testing methods. These methods will differ only in Step 1 of the general closed testing framework, that is, in the choice of the test statistic for H_{123}. For H_{12}, H_{13}, and H_{23} we will always simply use the test that rejects when $S_{ij} \geq c_\alpha$, that is, Step 2 of the closed testing framework is identical for all procedures. The four procedures have the following Step 1, with test statistics chosen so as to maximize power of the primary hypotheses:

(A) *Classic closed testing*: H_{123} is tested with test statistic S_{123};
(B) *Closed Tukey*: H_{123} is tested with test statistic $\tilde{S}_{123} = \max(S_{12}, S_{13}, S_{23})$;
(C) *Closed Dunnett*: H_{123} is tested with test statistic $\tilde{S}_1 = \max(S_{12}, S_{13})$;
(D) *Gatekeeping*: H_{123} is tested with test statistic S_{12}.

Unlike methods A and B, the methods C and D are not unique, since there is an analogous closed Dunnett method with test statistic $\tilde{S}_2 = \max(S_{12}, S_{23})$, taking μ_2 as the reference, and another one with $\tilde{S}_3 = \max(S_{13}, S_{23})$; Gatekeeping similarly has two additional variants. This makes for 8, rather than 4, methods in total. The variants of C and D differ only in the indexing of the hypotheses and are not fundamentally different.

Some of the procedures we have just proposed are special cases of previously proposed multigroup procedures. We do not claim novelty for any of them, since they are simple and direct consequences of the closed testing principle. Method A is a the three-group realization of the procedure of Shaffer (1979). Method B is the procedure of Student–Newman–Keuls (Fisher 1935; Newman 1939; Keuls 1952), which builds upon the proposal of Fisher (1935) to use of Student's t tests following an ANOVA F test, referred to as least significant difference (LSD). However, Fisher's LSD and Student–Newman–Keuls were abandoned or modified in subsequent years because they do not control the FWER in the general multiple group case, even though they do control in the special case of three groups (Hartley 1955; Hayter 1986). Method C was advocated by Marcus, Peritz, and Gabriel (1976), with variants by Finner (1990) and Hothorn (2020). We did not find a prior reference for method D, but we call it Gatekeeping because it is similar in spirit to the procedures of Dmitrienko, Tamhane, and Wiens (2008). In all cases the three-group setting is special: having only three groups simplifies a complex multi-group procedure (A and C) or rescues the validity of a method (B).

The proposed procedures are direct uniform improvements of the standard procedures described in Section 5, and should therefore always replace these procedures in applications. In particular, the closed testing procedures tend to reject more of the pairwise hypotheses after at least one rejection has been made. Procedure A is a uniform improvement of both the post hoc Tukey and Dunnett procedures. These latter procedures have the same Step 1, but use the more stringent critical value of $\tilde{c}_\alpha$ and $\tilde{c}^1_\alpha$, respectively, instead of c_α for H_{12}, H_{13} (and H_{23}) in Step 2. Unlike post hoc Dunnett, Procedure A also has a possibility for rejecting the secondary hypothesis H_{23}. Procedure B is uniformly more powerful than Tukey's HSD: both Closed Tukey and Tukey's HSD reject any of H_{12}, H_{13}, or H_{23} for which the test statistic exceeds $\tilde{c}_\alpha$, but once at least one hypothesis has been rejected in this way, Procedure B tests the remaining ones again at the reduced critical value c_α. Procedure C compares to Dunnett's method in the same way, while also adding the possibility that H_{23} may be rejected (Shaffer 1977). Procedure D does not uniformly improve one of the standard procedures. For this latter procedure, we note that the proper critical value for test statistic S_{12} in a test for H_{123} is simply c_α: since H_{123} is a subset of H_{12} (see Figure 1), any valid test for H_{12} is automatically a valid test for H_{123}.

Table 1 gives the probability that the closed testing methods reject at least one hypothesis more than the methods they uniformly improve. As a uniform improvement, the probability that it rejects fewer hypotheses is zero. The data are generated under a standard one-way ANOVA model with unknown $\sigma^2 = 1$ and $n = 6$ per group. We see from the table that the probability of improvement is substantial. The probabilities for the Dunnett methods are especially high, because the probability that the classical Dunnett methods rejects H_{23} is zero, while the corresponding probability may be large for its closed testing improvement. If H_{23} is disregarded, then probabilities become in the same order of magnitude for the Dunnett comparisons as for Tukey (data not shown).

7. Power of the Four Procedures

When to prefer which of the four procedures from the previous section? Obviously, we would like to maximize the probability of rejection of the hypotheses, prioritizing the primary hypotheses

Table 1. Probability that each closed testing procedure rejects at least one hypothesis more than the method it uniformly improves. CCT stands for classic closed testing.

Comparison	(μ_1,μ_2,μ_3)					
	(2,0,1)	(2,1,0)	(1,2,0)	(2,0,2)	(2,2,0)	(0,2,2)
CCT vs. post hoc Tukey	0.26	0.26	0.25	0.16	0.17	0.16
closed Tukey vs. Tukey's HSD	0.25	0.24	0.24	0.15	0.15	0.15
CCT vs. post hoc Dunnett	0.43	0.42	0.79	0.87	0.86	0.13
closed Dunnett vs. Dunnett	0.42	0.42	0.49	0.79	0.78	0.13

Table 2. Probability P(E) that the multiple comparisons procedure rejects fewer primary hypotheses than unadjusted testing, for Scenario D (H_{12} is primary).

Method	Statistic	(μ_1,μ_2,μ_3)					
		(2,0,1)	(2,1,0)	(1,2,0)	(2,0,2)	(2,2,0)	(0,2,2)
A	S_{123}	0.10	0.01	0.01	0.03	0.00	0.03
B	$\tilde{S}_{123}$	0.09	0.01	0.01	0.04	0.00	0.04
C	$\tilde{S}_1$	0.06	0.01	0.07	0.06	0.00	0.02
	$\tilde{S}_2$	0.06	0.08	0.01	0.02	0.00	0.06
	$\tilde{S}_3$	0.40	0.02	0.02	0.11	0.00	0.12
D	S_{12}	0.00	0.00	0.00	0.00	0.00	0.00
	S_{13}	0.53	0.01	0.27	0.85	0.01	0.07
	S_{23}	0.54	0.28	0.01	0.06	0.01	0.85

as argued in Section 4. We will argue in this Section that each of the four procedures A, B, C, and D is, generally, the preferred procedure for its corresponding Scenario A, B, C, and D.

Among many possible definitions of power in multiple testing (Senn and Bretz 2007; Gou et al. 2014), we analyze the power of the four procedures by comparing them to unadjusted testing, that is, testing without any multiple testing correction. Let us focus on Scenarios B, C, and D first, in which all primary hypotheses are pairwise hypotheses. We note that in these scenarios the closed testing procedure can never reject more of the primary hypotheses than unadjusted testing would, although it could reject fewer. We will analyze the event E that the multiple comparisons procedure rejects fewer primary hypotheses than unadjusted testing would. We try to minimize the probability of the event E.

We analyze the methods in the four scenarios in reverse order, starting with Scenario D. For all scenarios we will calculate P(E) for six configurations of (μ_1,μ_2,μ_3), and for all eight closed testing methods, including the variants of Methods C and D. The data are generated under a standard one-way ANOVA model with unknown $\sigma^2 = 1$ and $n = 6$ per group.

Table 2 gives P(E) for Scenario D. For this scenario there is an obvious winner that has P(E) $= 0$ whatever μ_1,μ_2,μ_3. The intuition for this case is simple enough: if only H_{12} is primary, we can focus the multiple testing procedure on that hypothesis by testing H_{123} and H_{12} with the same test, which implies a gatekeeping procedure that prioritizes the single primary hypothesis. Looking at the performance of the other closed testing procedures, we see from Table 2 that, while all other multiple comparisons procedures have some power loss compared to unadjusted testing, this power loss is small when the procedure used a test for H_{123} that involves a strong contribution of S_{ij} for the i,j with the largest difference $|\mu_i - \mu_j|$. For example, if $\mu_1 - \mu_3$ is largest, then the procedures that tests H_{123} with S_{13} or with $\tilde{S}_1 = \max(S_{12}, S_{13})$ have nearly the same power as the optimal procedure.

Table 3. Probability P(E) that the multiple comparisons procedure rejects fewer primary hypotheses than unadjusted testing, for Scenario C (H_{12} and H_{13} are primary).

Method	Statistic	(μ_1,μ_2,μ_3)					
		(2,0,1)	(2,1,0)	(1,2,0)	(2,0,2)	(2,2,0)	(0,2,2)
A	S_{123}	0.10	0.10	0.02	0.03	0.03	0.06
B	$\tilde{S}_{123}$	0.10	0.10	0.02	0.04	0.04	0.07
C	$\tilde{S}_1$	0.06	0.07	0.14	0.07	0.07	0.04
	$\tilde{S}_2$	0.06	0.41	0.02	0.02	0.11	0.12
	$\tilde{S}_3$	0.41	0.07	0.02	0.11	0.02	0.13
D	S_{12}	0.01	0.53	0.28	0.01	0.85	0.07
	S_{13}	0.53	0.01	0.27	0.85	0.01	0.07
	S_{23}	0.54	0.54	0.01	0.06	0.07	0.91

Table 4. Probability P(E) that the multiple comparisons procedure rejects fewer primary hypotheses than unadjusted testing, for Scenario B (all three pairwise hypotheses are primary).

Method	Statistic	(μ_1,μ_2,μ_3)					
		(2,0,1)	(2,1,0)	(1,2,0)	(2,0,2)	(2,2,0)	(0,2,2)
A	S_{123}	0.11	0.11	0.11	0.06	0.06	0.06
B	$\tilde{S}_{123}$	0.10	0.11	0.11	0.07	0.07	0.07
C	$\tilde{S}_1$	0.07	0.07	0.41	0.13	0.13	0.04
	$\tilde{S}_2$	0.07	0.41	0.08	0.04	0.13	0.12
	$\tilde{S}_3$	0.41	0.07	0.08	0.12	0.04	0.13
D	S_{12}	0.01	0.54	0.55	0.07	0.91	0.07
	S_{13}	0.54	0.02	0.54	0.91	0.07	0.07
	S_{23}	0.54	0.54	0.01	0.06	0.06	0.91

A similar but slightly less clear-cut picture emerges from Table 3, that covers Scenario C, in which H_{12} and H_{13} are primary. Here, none of the procedures has P(E) $= 0$ exactly, and there is no overall winner. However, we see that methods have small P(E) if they emphasize both the large differences and the primary hypotheses. The closed Dunnett procedure is the preferred procedure if the mean of the reference group is at the extreme. If the reference category is in the middle, closed Tukey or classic closed testing may do better. A gatekeeping procedure that emphasizes the largest difference can have good power even if that difference does not correspond to a primary hypothesis.

In Scenario B, shown in Table 4, there is again good power to be had if there is a priori knowledge which differences are largest: a gatekeeping or closed Dunnett that prioritize the test for which the true difference is largest, wins out. However, in absence of such knowledge, these methods can be risky, since they will do very badly if they happen to emphasize the small differences. Closed Tukey and classical closed testing are a safer choice with low P(E) overall.

Scenario A is different from the other three, since it is essentially a comparison of three different tests for the same hypothesis H_{123}. We can put it into the same framework as in the other scenarios, comparing P(E), if we view the ANOVA test as the standard test for H_{123}, and we look for maximal consistency between the multiple comparisons procedure and unadjusted testing. In this case Scenario A is much like Scenario D, with a single optimal procedure, classical closed testing, that achieves P(E) $= 0$. If alignment to the standard test of H_{123} is not important, we find ourselves in a classical situation of comparing power for different tests of the same hypothesis H_{123}. Table 5 gives the power of the implied tests. As in the other scenarios, we see that the best power can be had from a test that focuses

Table 5. Power for rejecting H_{123}.

Method	Statistic	(μ_1, μ_2, μ_3)					
		(2,0,1)	(2,1,0)	(1,2,0)	(2,0,2)	(2,2,0)	(0,2,2)
A	S_{123}	0.80	0.80	0.80	0.90	0.90	0.90
B	$\tilde{S}_{123}$	0.81	0.80	0.80	0.90	0.89	0.89
C	$\tilde{S}_1$	0.84	0.83	0.50	0.83	0.83	0.92
	$\tilde{S}_2$	0.84	0.50	0.83	0.92	0.83	0.83
	$\tilde{S}_3$	0.50	0.83	0.83	0.84	0.92	0.83
D	S_{12}	0.90	0.37	0.36	0.89	0.05	0.89
	S_{13}	0.37	0.89	0.37	0.05	0.89	0.89
	S_{23}	0.37	0.37	0.90	0.90	0.90	0.05

power on the largest true difference, but, in the absence of such knowledge, the test using S_{123} or $\tilde{S}_{123}$ are a low-risk option.

To summarize, to minimize the probability that we lose out on some rejections compared to unadjusted testing, we must consider both the distinction between primary and secondary hypotheses, and any a priori knowledge on the values of μ_1, μ_2 and μ_3. In the situation that the primary hypotheses are also the hypotheses for which we expect the largest differences $|\mu_i - \mu_j|$, the situation is clear-cut: with scenario A, B, C, or D we should prefer the corresponding method A, B, C, or D.

If the hypotheses of most interest do not necessarily correspond to the hypotheses for which we expect largest effect size, the choice of method becomes more subtle. If there is a reliable a priori idea where the true values of μ_1, μ_2, and μ_3 could be, this could guide the choice of method. For example, if $|\mu_1 - \mu_2|$ is strongly expected to be the largest true difference, a we can expect gatekeeping based on S_{12} to have good power; if $\mu_2 \approx \mu_3$ is expected, then closed Dunnett using $\tilde{S}_1$ is preferable.

In the more common situation that we are not willing to gamble on a priori guesses about the means, however, the rule of thumb remains that it is a relatively safe choice in every scenario to choose the corresponding method. A possible exception is Scenario C, in which classical closed testing or closed Tukey could be preferred to closed Dunnett if researchers also want good power in the situation that the reference category could be the middle one.

8. Paradoxical Outcomes

The logical relationships between the hypotheses, displayed in Figure 1, dictate that the number of true hypotheses may be 0, 1 or 4, but never 2 or 3. Consequently, the number of false hypotheses should be 0, 3, or 4. The result of the test procedure, however, may not always conform to this.

The most well-known of these situations is the frustrating event that H_{123} is rejected, but none of H_{12}, H_{13}, or H_{23} (Gabriel 1969; Romano, Shaikh, and Wolf 2011). In this case, we may claim that at least two more hypotheses are false, but that we are not confident which ones. This event occurs frequently with, for example, ANOVA followed by Bonferroni (Sedgwick 2014). Fortunately, this event less of an issue with closed testing in three groups: it is impossible with procedures B, C and D, due to logical implications between the test outcomes, and extremely rare with procedure A, since S_{123} and $\tilde{S}_{123}$ are highly correlated.

Another paradoxical outcome can occur when only one of H_{12}, H_{13}, H_{23} is rejected. In such cases, we may claim that at least one more hypothesis must be false, only we are not sure which one. This type of paradoxical outcome can occur with any of the above described methods, since it may also occur with unadjusted testing. However, closed testing procedures also reduce the probability of this event, compared to, for example, Tukey's HSD: they increase the probability of rejecting the second pairwise hypothesis after the first one is rejected.

9. Applying the Four Procedures: Adjusted *p*-Values

To use these four methods in the ANOVA context with standard statistical software it is easiest to work from adjusted p-values. Adjusted p-values can be calculated for any multiple testing procedure. They are defined for any hypothesis as the smallest FWER level α at which the hypothesis would be rejected. Therefore, a hypothesis is rejected by the multiple testing procedure if and only if its adjusted p-value is at most α (Rosenthal and Rubin 1983; Wright 1992).

For the one-way ANOVA design with equal groups, most statistical software packages return adjusted p-values $\tilde{p}_{12}^{\text{Tuk}}$, $\tilde{p}_{13}^{\text{Tuk}}$, $\tilde{p}_{23}^{\text{Tuk}}$, for H_{12}, H_{13}, and H_{23}, respectively, for Tukey's HSD method and $\tilde{p}_{12}^{\text{Dun}}$, $\tilde{p}_{13}^{\text{Dun}}$ for H_{12}, and H_{13} for Dunnett's method. From these, together with the unadjusted p-values p_{12}, p_{13}, p_{23}, and p_{123}, we can calculate adjusted p-values for all four methods we have introduced.

In the general closed testing framework, we reject each hypothesis if both the hypothesis itself and the implying hypothesis H_{123} have been rejected,. The adjusted p-value of H_{ij} in a closed testing procedure is therefore

$$\tilde{p}_{ij} = \max(p_{ij}, \tilde{p}_{123}),$$

where $\tilde{p}_{123}$ is the p-value for H_{123} in the procedure. These we can calculate for each of the four procedures as follows:

$$\begin{aligned}
\tilde{p}_{123}^{A} &= p_{123};\\
\tilde{p}_{123}^{B} &= \min(\tilde{p}_{12}^{\text{Tuk}}, \tilde{p}_{13}^{\text{Tuk}}, \tilde{p}_{23}^{\text{Tuk}});\\
\tilde{p}_{123}^{C} &= \min(\tilde{p}_{12}^{\text{Dun}}, \tilde{p}_{13}^{\text{Dun}});\\
\tilde{p}_{123}^{D} &= p_{12}.
\end{aligned}$$

To understand these expressions for Procedures B and C, remark that we reject H_{123} there if the Tukey's HSD and Dunnett procedures, respectively, reject at least one hypothesis, which happens when the Tukey's HSD- (or Dunnett-) adjusted p-value for at least one hypothesis is less than α. Using these formulae we can apply these four closed testing procedures with any software that can apply the usual Tukey's HSD and Dunnett procedures. The adjusted p-value for H_{123} in all four procedures is simply $\tilde{p}_{123}^{A}$, $\tilde{p}_{123}^{B}$, $\tilde{p}_{123}^{C}$, or $\tilde{p}_{123}^{D}$.

10. Three Groups beyond ANOVA

Three-group comparisons occur in many more contexts than ANOVA, for example when comparing three proportions using chi-squared tests in a 2×3 table, when performing nonparametric analysis with Kruskal–Wallis tests, when comparing three survival curves using a log-rank test, or in regression models when considering a categorical covariate with three levels. In all such cases, we can formulate a global null hypothesis

H_{123} of equality of all three groups and corresponding pairwise hypotheses H_{12}, H_{13}, and H_{23}. Regardless of the model considered, the logical relationships between hypotheses and the distinction between primary and secondary hypotheses remain the same, so that the same four scenarios arise.

We can also define four methods for these four scenarios in more or less the same way. Practically, however, there is a difference between the ANOVA model context and other models. While classical closed testing and gatekeeping (A and D) may still be used by simply applying these methods to p-values from model-appropriate tests (e.g., likelihood ratio or Wald tests) analogues of Dunnett's and Tukey's HSD methods are generally unavailable in commercial statistical software packages. Asymptotic versions of these methods can be used if the estimates of the three parameters are asymptotically normal, using the *multcomp* package in R (Hothorn, Bretz, and Westfall 2008), and we illustrate how in Section 12. The *multcomp* package can also be used if different contrasts are of interest, for example, $\mu_1 - (\mu_2 + \mu_3)/2$ (Dunnett and Tamhane 1992). In case R is out of reach of practitioners, a closed testing method may be constructed using any valid and powerful test for H_{123} that the software offers. Sometimes, only the tests of methods A and D will be available; in that case method A is recommended in Scenarios B and C, since it does not require an arbitrary choice of a primary hypothesis.

11. ANOVA Example

We illustrate the four procedures with an one-way ANOVA example from Dobson's (1983) book. Suppose that genetically similar seeds are randomly assigned to be raised either under standard conditions (control) or in two different nutritionally enriched environments (Treatments I and II). After a predetermined period all plants are harvested, dried and weighed. The results, expressed as dried weight in grams, for samples of $n = 10$ plants from each group are given in Table 6.

The response, plant weight, depends on one factor, growing condition, with three levels—control, treatment I and treatment II. The choice of the hypotheses of primary interest depends on the context. Scenario A is appropriate if we would first and foremost want to show that there is some effect of different growing conditions, regardless of which. Scenario B would be chosen if we would be equally interested in showing a difference between any of the groups, but if only rejecting the global hypothesis would be unsatisfactory. Scenario C would be appropriate if we would be primarily interested in finding at least one of the treatments is different from the control. Scenario D prioritizing H_{12} would be most appropriate if demonstrating the effectiveness of treatment I with respect to the control would be of primary interest.

The estimated means are $\hat{\mu}_1 = 5.03$, $\hat{\mu}_2 = 4.66$ and $\hat{\mu}_3 = 5.53$, with estimated variance of $\hat{\sigma}^2 = 0.389$. The unadjusted

Table 6. Dried weights of plants grown from under three different conditions (data from Dobson (1983), Table 7.1).

Group 1 (control)	4.17	5.58	5.18	6.11	4.50	4.61	5.17	4.53	5.33	5.14
Group 2 (treatment I)	4.81	4.17	4.41	3.59	5.87	3.83	6.03	4.89	4.32	4.69
Group 3 (treatment II)	6.31	5.12	5.54	5.50	5.37	5.29	4.92	6.15	5.80	5.26

Table 7. Adjusted p-values for the four hypotheses and four methods in the plant growth example (left) and for the alternative analysis based on permutation tests (right).

	F-tests				*Permutation tests*			
Method	H_{12}	H_{13}	H_{23}	H_{123}	H_{12}	H_{13}	H_{23}	H_{123}
(A) Classic closed testing	0.194	0.088	0.016	0.016	0.247	0.048	0.017	0.017
(B) Closed Tukey	0.194	0.088	0.012	0.012	0.247	0.048	0.012	0.012
(C) Closed Dunnett	0.194	0.153	0.153	0.153	0.247	0.205	0.205	0.205
(D) Gatekeeping	0.194	0.194	0.194	0.194	0.247	0.247	0.247	0.247

p-values for standardized group differences are $p_{12} = 0.194$, $p_{13} = 0.088$, $p_{23} = 0.004$. Adjusted p-values for Tukey's HSD and Dunnett methods are $\tilde{p}_{12}^{\text{Tuk}} = 0.391$, $\tilde{p}_{13}^{\text{Tuk}} = 0.198$, $\tilde{p}_{23}^{\text{Tuk}} = 0.012$ and $\tilde{p}_{12}^{\text{Dun}} = 0.323$, $\tilde{p}_{13}^{\text{Dun}} = 0.153$, respectively, giving $\tilde{p}_{123}^{B} = 0.012$ and $\tilde{p}_{123}^{C} = 0.153$. Together with $\tilde{p}_{123}^{A} = 0.016$ for the ANOVA F test, we obtain the adjusted p-values displayed in Table 7 for the four hypotheses and the four methods.

We see that at the significance level of $\alpha = 5\%$, closed Dunnett and gatekeeping do not reject any hypothesis, while classic closed testing and closed Tukey reject H_{123} and H_{23}. We note that this result is logically incomplete: if H_{23} is false, then we know that at least one of H_{12} or H_{13} must be false, but we cannot confidently say which one.

Alternative Analyses

The methods A, B, C, and D are not tied to F-tests, and naturally generalize to other tests. We give some examples below. Of course, for any of these methods to be valid, the researcher should choose the test procedure independently of the data.

If we believe that variances may differ between groups when the means do, we would prefer a two-sample t-test over the F-test for the pairwise hypotheses. This results in $p_{12} = 0.250$, $p_{13} = 0.048$ and $p_{23} = 0.009$, which gives the rejection of H_{13} and H_{23} at $\alpha = 5\%$ for methods A and B, and the paradoxical outcome disappears. However, if we replace S_{ij} by two-sample t-tests, the rejection of H_{123} by $\tilde{S}$ or $\tilde{S}_1$ no longer guarantees that at least one H_{ij} will be rejected by t-test. The four methods with t-tests represent further variants of closed testing, with method A becoming Fisher's LSD.

Randomization in the experimental design leads naturally to the use of permutation tests (Ludbrook and Dudley 1998), and we could also use methods A, B, C and D in a permutation framework. A standard choice is to consider non-standardized test statistics $T_{ij} = (\hat{\mu}_i - \hat{\mu}_j)^2$ for testing H_{ij}, and $T_{123} = T_{12} + T_{23} + T_{13}$, $\tilde{T} = \max(T_{12}, T_{13}, T_{23})$, $\tilde{T}_1 = \max(T_{12}, T_{13})$, and T_{12} for testing H_{123} in methods A, B, C, and D, respectively. The construction of the permutation null distribution under H_{123} proceeds as follows. The observations of the groups to be compared are pooled, and the test statistic is recalculated for every permutation of the group labels. Then, the permutation p-value is calculated as the proportion of permutations where the test statistic is greater than or equal to the value computed on the original data. If the cardinality of the set of all possible permutations is too large, then one may use a subset of randomly chosen elements (Hemerik and Goeman 2018). Note that, while tests of H_{123} uses a global permutation distribution, constructed by permuting the observations of all three groups, tests of H_{ij}

use a local permutation distribution, constructed by permuting the observations of groups i and j (Petrondas and Gabriel 1983). For the data of this section, based on 10^6 random permutations, we obtain $p_{12} = 0.247$, $p_{13} = 0.048$ and $p_{23} = 0.008$, $\tilde{p}^A_{123} = 0.017$, $\tilde{p}^B_{123} = 0.012$ and $\tilde{p}^C_{123} = 0.205$, resulting in the rejection of H_{123}, H_{13} and H_{23} at $\alpha = 5\%$ for methods A and B. Adjusted p-values are reported in Table 7 for comparison with the main analysis. Rank tests are a special case of permutation tests, replacing the observations with their ranks. We obtain $p_{12} = 0.197$, $p_{13} = 0.063$ and $p_{23} = 0.009$ from Wilcoxon–Mann–Whitney tests, $\tilde{p}^A_{123} = 0.014$ from Kruskal–Wallis test, $\tilde{p}^B_{123} = 0.010$ and $\tilde{p}^C_{123} = 0.172$. With rank tests, methods A and B reject H_{123} and H_{23} at $\alpha = 5\%$. R code for reproducing the results of this Section and the next are available in Supplementary Material C.

12. ANCOVA Example

The methods we have described for the ANOVA setting generalize to all statistical models in which there are thee parameters to compare. The methods A, B, C, and D are applicable in all such models as long as we can rely on (asymptotic) normality of the parameters, since we can use the *multcomp* package to find the distribution of the test statistics. We will illustrate this with an ANCOVA model.

As with ANOVA, we are interested in comparing means for groups while controlling for the effects of other covariates that are not of primary interest. Table 8 displays data from Winer (1971), discussed in Dobson (1983). The response y is the achievement score, the levels $i = 1, 2$ and 3 of the group factor represent three different training methods, and the covariate x is the aptitude score measured before training commenced. We want to compare the training methods, taking into account differences in initial aptitude between the three groups of subjects.

We assume that the response in group i is normally distributed with mean $\mu_i(x)$ and variance σ^2, with

$$\mu_i(x) = \gamma + \tau_i + \beta(x - \bar{x})$$

where γ is the common mean, τ_i is the ith group effect such that $\sum_i \tau_i = 0$, β is the regression slope and $\bar{x}$ is the average covariate value.

Analysis of covariance compares the adjusted means $\hat{\mu}_i(\bar{x}_i)$, that is, the estimated group means adjusted for the group average covariate values, which are equal to $\hat{\mu}_1(\bar{x}_1) = 4.89$, $\hat{\mu}_2(\bar{x}_2) = 7.08$ and $\hat{\mu}_3(\bar{x}_3) = 6.75$ by least-square estimation. Let Y and X denote the response vector and the design matrix, respectively, and let $\hat{\theta} = (X'X)^{-1}XY$ be the least-square estimator. For the 3×4 matrix K corresponding to pairwise contrasts, we have

$$K\hat{\theta} = \begin{pmatrix} \hat{\mu}_2(\bar{x}_2) - \hat{\mu}_1(\bar{x}_1) \\ \hat{\mu}_3(\bar{x}_3) - \hat{\mu}_1(\bar{x}_1) \\ \hat{\mu}_3(\bar{x}_3) - \hat{\mu}_2(\bar{x}_2) \end{pmatrix} \overset{H_{123}}{\sim} N(0, \Sigma),$$

with $\Sigma = \sigma^2 K(X'X)^{-1}K'$, and the standardized vector of test statistics $T = D^{-1/2}K\hat{\theta}$ follows a multivariate t distribution with $\nu = 3n - 4$ degrees of freedom and correlation matrix $R = D^{-1/2}\hat{\Sigma}D^{-1/2}$, with $D = \text{diag}(\hat{\Sigma})$ and $\hat{\Sigma} = \hat{\sigma}^2 K(X'X)^{-1}K'$.

Unadjusted p-values $p_{12} = 0.0002$, $p_{13} = 0.0004$ and $p_{23} = 0.4563$ are obtained from the Student t distribution or, equivalently, from partial F tests of H_{ij} comparing the constrained model with $\tau_i = \tau_j$ to the unconstrained one. Tukey's HSD adjusted p-values $\tilde{p}^{\text{Tuk}}_{12} = 0.0004$, $\tilde{p}^{\text{Tuk}}_{13} = 0.0011$, $\tilde{p}^{\text{Tuk}}_{23} = 0.7302$ are obtained by calculating the distribution of the maximum T (Genz and Bretz 2002). Dunnett's p-values $\tilde{p}^{\text{Dun}}_{12} = 0.0003$ and $\tilde{p}^{\text{Dun}}_{13} = 0.0008$ can be obtained in a similar fashion. This gives $\tilde{p}^B_{123} = 0.0004$ and $\tilde{p}^C_{123} = 0.0003$ for methods B and C, and the ANCOVA F-test gives $\tilde{p}^A_{123} = 0.0002$ for method A. In this example all the methods reject H_{123}, H_{12}, and H_{13} at the significance level of 5%.

Table 8. Achievement scores (data from Winer 1971, p. 766)

	Training method i					
	1		2		3	
Unit u	$y_{u,1}$	$x_{u,1}$	$y_{u,2}$	$x_{u,2}$	$y_{u,3}$	$x_{u,3}$
1	6	3	8	4	6	3
2	4	1	9	5	7	2
3	5	3	7	5	7	2
4	3	1	9	4	7	3
5	4	2	8	3	8	4
6	3	1	5	1	5	1
7	6	4	7	2	7	4

13. Discussion

We have presented four closed testing-based methods for pairwise comparisons between three parameters. These four methods are tailored to four scenarios, distinguished by the a priori choice which of the hypotheses are primary and which are secondary. The procedures we have described are tailored to the three-group problem, and are uniformly more powerful than several frequently used procedures.

Closed testing procedures are complex in the general multigroup case, but remain relatively simple in the important three-group case. The four methods can be applied for the ANOVA with any statistial software that provides Tukey's HSD and Dunnett's procedures, or in general models with the *multcomp* package in R.

The three-group situation is special. We argue that statistics textbooks for practitioners should not jump immediately from the two-group to the multigroup case, but should consider the three group case explicitly. If nothing else, then they should say that the need for correcting post hoc tests after a significant ANOVA arises only with four or more groups (Method A). Only if researchers are interested in the case of four or more groups, or if simultaneous confidence intervals of the group means are needed, do they need to consult specialized multiple comparisons literature (e.g., Spurrier and Isham 1985).

It cannot be emphasized enough that in all multiple comparisons procedures the choice of method, that is, the precise tests to use for all of H_{123}, H_{12}, H_{13}, and H_{23}, should be made from subject-matter considerations, independently of the data. Choosing as primary hypotheses the hypotheses with largest $|\hat{\mu}_i - \hat{\mu}_j|$ is sure to lead to an inflated false positive rate (Simmons, Nelson, and Simonsohn 2011).

Supplementary Material

A: Derivation and extension: Derivation of equation (1) and extension to the four-group case. (pdf).
B: Simulation R-code: R-code for the simulations in Sections 3, 6 and 7. (.R file).
C: Examples R code: R-code for the applications in Sections 11, 12 and for the four-group example in Supplementary Material A. (.Rmd file).

References

Begun, J. M., and Gabriel, K. R. (1981), "Closure of the Newman–Keuls Multiple Comparisons Procedure," *Journal of the American Statistical Association*, 76, 241–245. [168]

Benjamini, Y. (2019), "Selective Inference: The Silent Killer of Replicability," in *The Henry L. Rietz Lecture, ASA Joint Statistical Meetings*, Denver, Colorado. [169]

Bergmann, B., and Hommel, G. (1988), "Improvements of General Multiple Test Procedures for Redundant Systems of Hypotheses," in *Multiple Hypothesenprüfung/Multiple Hypotheses Testing*, eds. P. Bauer, G. Hommel, and E. Sonnemann, Berlin: Springer, pp. 100–115. [168]

Bretz, F., Hothorn, T., and Westfall, P. (2011), *Multiple Comparisons Using R*, Boca Raton, FL: CRC Press. [168]

Bretz, F., Maurer, W., Brannath, W., and Posch, M. (2009), "A Graphical Approach to Sequentially Rejective Multiple Test Procedures," *Statistics in Medicine*, 28, 586–604. [169]

Burman, C.-F., Sonesson, C., and Guilbaud, O. (2009), "A Recycling Framework for the Construction of Bonferroni-Based Multiple Tests," *Statistics in Medicine*, 28, 739–761. [169]

Cui, X., Dickhaus, T., Ding, Y., and Hsu, J. C. (2021), *Handbook of Multiple Comparisons*, Boca Raton, FL: CRC Press. [168]

Dickhaus, T. (2014). *Simultaneous Statistical Inference*, Berlin: Springer. [168]

Dmitrienko, A., Tamhane, A. C., and Wiens, B. L. (2008), "General Multistage Gatekeeping Procedures," *Biometrical Journal*, 50, 667–677. [171]

Dobson, A. J. (1983), *Introduction to Statistical Modelling*, New York: Springer. [174,175]

Dunnett, C. W. (1955), "A Multiple Comparison Procedure for Comparing Several Treatments With a Control," *Journal of the American Statistical Association*, 50, 1096–1121. [170]

Dunnett, C. W., and A. C. Tamhane (1992), "A Step-Up Multiple Test Procedure," *Journal of the American Statistical Association,* 87, 162–170. [174]

Field, A. (2013), *Discovering Statistics Using IBM SPSS Statistics* (4th ed.), Sage. [168]

Finner, H. (1990), "On the Modified s-Method and Directional Errors," *Communications in Statistics–Theory and Methods*, 19, 41–53. [171]

Fisher, R. A. (1935), *The Design of Experiments,* Edinburgh: Oliver & Boyd. [171]

Gabriel, K. R. (1969), "Simultaneous Test Procedures–Some Theory of Multiple Comparisons," *The Annals of Mathematical Statistics*, 40, 224–250. [173]

Glover, T., and Mitchell, K. (2008), *An Introduction to Biostatistics*, Long Grove, IL: Waveland Press. [168]

Goeman, J. J., Hemerik, J., and Solari, A. (2021), "Only Closed Testing Procedures Are Admissible for Controlling False Discovery Proportions," *The Annals of Statistics*, 49, 1218–1238. [168,170]

Gou, J., Tamhane, A. C., Xi, D., and Rom, D. (2014), "A Class of Improved Hybrid Hochberg-Hommel Type Step-Up Multiple Test Procedures," *Biometrika*, 101, 899–911. [172]

Hartley, H. (1955), "Some Recent Developments in Analysis of Variance," *Communications on Pure and Applied Mathematics*, 8, 47–72. [171]

Hayter, A. J. (1986), "The Maximum Familywise Error Rate of Fisher's Least Significant Difference Test," *Journal of the American Statistical Association*, 81, 1000–1004. [171]

Hemerik, J., and Goeman, J. (2018), "Exact Testing With Random Permutations," *Test*, 27, 811–825. [174]

Hochberg, Y., and Tamhane, A. C. (1987), *Multiple Comparison Procedures*, Hoboken, NJ: Wiley. [168]

Hothorn, L. A. (2020), "Comparisons of Multiple Treatment Groups With a Negative Control or Placebo Group: Dunnett Test vs. Closed Test Procedure," arXiv: 2012.04277. [171]

Hothorn, T., Bretz, F., and Westfall, P. (2008), "Simultaneous Inference in General Parametric Models," *Biometrical Journal*, 50, 346–363. [174]

Hsu, J. (1996), *Multiple Comparisons: Theory and Methods*, Boca Raton, FL: CRC Press. [168]

Keuls, M. (1952), "The Use of the "Studentized Range" in Connection With an Analysis of Variance," *Euphytica*, 1, 112–122. [171]

Ludbrook, J., and Dudley, H. (1998), "Why Permutation Tests are Superior to t and f Tests in Biomedical Research," *The American Statistician*, 52, 127–132. [174]

Marcus, R., E. Peritz, and K. R. Gabriel (1976), "On Closed Testing Procedures With Special Reference to Ordered Analysis of Variance," *Biometrika*, 63, 655–660. [168,170,171]

Miller, R. G. (1981), *Simultaneous Statistical Inference* (2nd ed.), Springer. [168]

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. [171]

Petrondas, D. A., and Gabriel, K. R. (1983), "Multiple Comparisons by Rerandomization Tests," *Journal of the American Statistical Association*, 78, 949–957. [175]

Rom, D. M., and Holland, B. (1995), "A New Closed Multiple Testing Procedure for Hierarchical Families of Hypotheses," *Journal of Statistical Planning and Inference*, 46, 265–275. [168]

Romano, J. P., Shaikh, A., and Wolf, M. (2011), "Consonance and the Closure Method in Multiple Testing," *The International Journal of Biostatistics*, 7, 1–25. [173]

Rosenthal, R., and Rubin, D. B. (1983), "Ensemble-Adjusted p Values," *Psychological Bulletin*, 94, 540. [173]

Sedgwick, P. (2014), "One Way Analysis of Variance: Post Hoc Testing," *British Medical Journal*, 349, g7067. [173]

Senn, S., and Bretz, F. (2007), "Power and Sample Size When Multiple Endpoints are Considered," *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry*, 6, 161–170. [172]

Shaffer, J. P. (1977), "Multiple Comparisons Emphasizing Selected Contrasts: An Extension and Generalization of Dunnett's Procedure," *Biometrics*, 33, 293–303. [171]

Shaffer, J. P. (1979), "Comparison of Means: An F Test Followed by a Modified Multiple Range Procedure," *Journal of Educational Statistics*, 4, 14–23. [171]

——— (1986), "Modified Sequentially Rejective Multiple Test Procedures," *Journal of the American Statistical Association*, 81, 826–831. [169]

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011), "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant," *Psychological Science*, 22, 1359–1366. [175]

Sonnemann, E. (2008), "General Solutions to Multiple Testing Problems," *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 50, 641–656. [170]

Spurrier, J. D., and Isham, S. P. (1985), "Exact Simultaneous Confidence Intervals for Pairwise Comparisons of Three Normal Means," *Journal of the American Statistical Association*, 80, 438–442. [175]

Stevens, J. P. (2013), *Intermediate Statistics: A Modern Approach*, Milton Park: Routledge. [168]

Tabachnick, B. G., Fidell, L. S., and Ullman, J. B. (2007), *Using Multivariate Statistics*, Vol. 5, London: Pearson. [168]

Tippett, L. H. (1925), "On the Extreme Individuals and the Range of Samples Taken From a Normal Population," *Biometrika*, 17, 364–387. [170]

Tukey, J. W. (1949), "Comparing Individual Means in the Analysis of Variance," *Biometrics*, 5, 99–114. [170]

——— (1953), "The Problem of Multiple Comparisons," in *The Collected Works of John W. Tukey VIII. Multiple Comparisons: 1948–1983*, ed. Braun, H. I., New York: Chapman and Hall, pp. 1–300. [169]

Winer, B. J. (1971), *Statistical Principles in Experimental Design* (2nd ed.). New York: McGraw-Hill. [175]

Wright, S. P. (1992), "Adjusted p-Values for Simultaneous Inference," *Biometrics*, 48, 1005–1013. [173]