For Post Hoc's Sake: Determining Sample Size for Tukey Multiple Comparisons in 4-Group ANOVA

Gordon P. Brooks Ohio University

Qian An Ohio University

Yanju Li Georgia State University

George A. Johanson Ohio University

The determination of an appropriate sample size is a difficult, yet critically important, element in the research design process. Sample sizes in ANOVA are most often based on an overall standardized difference in the means, but these recommended sample sizes provide statistical power only for the omnibus *F* test. Adequate sample size for the omnibus test does not necessarily provide sufficient statistical power for the post hoc multiple comparisons typically performed following a statistically significant omnibus *F* test. The purpose of this paper is to use Monte Carlo techniques to determine sample sizes that will provide adequate statistical power for Tukey post hoc multiple comparison procedure (MCP) in 4-group ANOVA. Sample size tables are included.

In the planning stage of a research study, investigators are often uncertain about the minimum number of subjects needed to adequately test a hypothesis of interest (Olejnik, 1984). Since Cohen proposed sample size determination in the planning, execution and interpretation of statistical analysis, researchers have increasingly realized the importance of sample size (Brewer & Sindelar, 1988). The number of cases that should be involved in a research study, particularly in hypothesis testing, has become a critical concern (Olejnik, 1984).

Most researchers generally know that the sample size determination is functionally related to the significance level, statistical power, and the effect size. The required sample size is inversely related to significance level, given that power and effect size are held constant. A priori power, the probability of rejecting a null hypothesis that is indeed false, will inform researchers how many subjects will be needed for adequate power (Light, Singer, & Willett, 1990). Effect size is the degree to which the null hypothesis is false (Olejnik, 1984). Different effect size values will require different sample sizes. In addition, Brooks and Johanson (2011) pointed out that the sample size needed for an adequate test of a hypothesis is affected by the particular statistical analysis strategy that researchers choose. For example, when ANOVA will be used, adequate sample size for the omnibus test does not necessarily provide adequate statistical power for the post hoc multiple comparisons typically performed in ANOVA.

This paper aims to demonstrate how to choose the sample size that will provide adequate power in the Tukey post hoc multiple comparison procedure (MCP). Pairwise multiple comparisons among four groups will be studied and results will be organized into recommended sample size tables for Tukey test.

Theoretical Perspectives

One of the more commonly used research methods is group comparisons, and analysis of variance (ANOVA) is one well-known approach to compare groups based on means (Wilcox, 2002). As of this writing, there were 34,000 citations for 'analysis of variance' since 2021 on Google Scholar. Barnette and McLean (1999) mentioned that most researchers follow the practice of conducting post hoc pairwise MCPs after a significant omnibus *F* test, but when a researcher encounters more than two comparisons, control of the Type I error becomes a concern. In fact, after Fisher developed the process of analysis of variance (ANOVA), he realized the potential problem of Type I error inflation when multiple *t* tests were conducted on three or more groups; accordingly, Fisher suggested that researchers should use a more stringent alpha based on his concern (Barnette & McLean, 1999). Klockars and Hancock (1998) proposed that MCPs are used to invoke control over the family-wise Type I error. In practice, there are several commonly used multiple comparison procedures, such as Dunn-Bonferroni, Dunn-Šidák, and Tukey's HSD. Tukey's HSD is most highly recommended as an unprotected test (Barnette & McLean, 1999) and therefore does not need to be protected by a statistically significant ANOVA.

The replication crisis has promoted preregistration of studies and a confirmative mindset (Nosek et al., 2018; Wagenmakers et al., 2012). As a consequence, it has cast a shadow on exploratory research and post-hoc analyses. However recent research (e.g., Rubin & Donkin, 2022) has moderated this position by closely examining advantages and disadvantages of both exploratory and confirmatory approaches. They contend that "…exploratory hypothesis tests can have more advantages and fewer disadvantages than confirmatory tests and that, consequently, exploratory hypothesis tests can yield more compelling research conclusions than confirmatory tests" (p. 21). We concur with this position.

Cohen (1988) illustrated sample size selection in ANOVA, in which the effect size and the power are related to the sample size determination. Levin (1975) proposed an approach for determining the sample size needed for ANOVA and for planned contrasts. However, most studies of ANOVA are limited to the omnibus *F* test, while omitting the appropriate sample size selection in the MCPs, which usually follows a significant *F* test.

Brooks and Johanson (2011) demonstrated the sample size selection for MCPs in ANOVA with three groups based on Tukey and Bonferroni tests, which clearly showed that adequate statistical power for the omnibus ANOVA *F* test does not guarantee enough statistical power for given pairwise comparisons performed post hoc. They also derived a pattern between the required sample sizes for the omnibus ANOVA and the sample sizes needed for the MCPs.

Overall, there have been few studies that focused on the sample size determination in MCP. In this context, this paper will focus on the sample size selection with four groups when using the Tukey MCP.

Methods and Data Source

The Monte Carlo program called *MC4G: Monte Carlo Analyses for up to 4 Groups* (Raffle & Brooks, 2005) has been provided by Brooks (2008) to perform Monte Carlo analyses for *t* tests and ANOVA in a Windows environment. The program uses Monte Carlo techniques to find sample sizes that meet a given statistical power level. We set up many conditions in MC4G and ran the simulations in this program. Conditions in this study included four-group one-way ANOVAs performed at a .05 level of significance. Desired statistical power was set at .80. There were 10,000 replications performed for the final sample size analysis in each condition. Condition values were entered into the graphical user interface in MC4G.

In all simulations, normally distributed standardized data was generated to fit the given conditions for each simulation. That is, all variances were set to 1.0, while group means varied from 0.0 to 0.8, depending on the given effect size. In particular, all possible patterns of mean differences (effect sizes) with at least a 0.2 standardized difference were analyzed in an effort to identify sample size relationships between the omnibus test and the Tukey MCP. For example, group means across the four groups of 0.0, 0.0, 0.2, and 0.4 result in the same effect sizes (mean differences) as means of 0.4, 0.4, 0.6, and 0.8—so the pattern of six possible differences among the means (i.e., 0.0, 0.2, 0.2, 0.2, 0.4, 0.4) will only be included once.

Results

Results are provided in Table 1 and Table 2. Totally, 45 patterns of standardized means among four groups were analyzed; then, the sample sizes for the omnibus ANOVA test and the multiple comparison procedure were listed. In each of these patterns, there were six possible mean differences (effect sizes) possible among the four group means. In addition, the relative efficiency, the omnibus per group sample size divided by multiple comparison per group sample size, was finally calculated. There were several interesting findings:

Table 1. Sample size results for the Tukey HSD Multiple Comparison Procedure for the primary Monte Carlo design at statistical power of .80

Group 1 mean	Group 2 mean	Group 3 mean	Group 4 mean	Effect Size	Comparison Tested	Total Sample Size	Sample Size per Group	Relative Efficiency [a]
0.0	0.0	0.0	0.2		Omnibus	1432	358	
				0.00	G1 v G2	*	*	
				0.00	G2 v G3	*	*	
				0.00	G3 v G1	*	*	
				0.20	G4 v G1	2320	580	1.62
				0.20	G4 v G2	2348	587	1.64
				0.20	G4 v G3	2316	579	1.62

0.0	0.0	0.0	0.3		Omnibus	636	159	
				0.00	G1 v G2	*	*	
				0.00	G2 v G3	*	*	
				0.00	G3 v G1	*	*	
				0.30	G4 v G1	1032	258	1.62
				0.30	G4 v G2	1056	264	1.66
				0.30	G4 v G3	1064	266	1.67
0.0	0.0	0.0	0.4		Omnibus	376	94	
				0.00	G1 v G2	*	*	
				0.00	G2 v G3	*	*	
				0.00	G3 v G1	*	*	
				0.40	G4 v G1	588	147	1.56
				0.40	G4 v G2	596	149	1.59
				0.40	G4 v G3	588	147	1.56
0.0	0.0	0.0	0.5		Omnibus	244	61	
				0.00	G1 v G2	*	*	
				0.00	G2 v G3	*	*	
				0.00	G3 v G1	*	*	
				0.50	G4 v G1	376	94	1.54
				0.50	G4 v G2	380	95	1.56
				0.50	G4 v G3	376	94	1.54
0.0	0.0	0.0	0.6		Omnibus	172	43	
				0.00	G1 v G2	*	*	
				0.00	G2 v G3	*	*	
				0.00	G3 v G1	*	*	
				0.60	G4 v G1	264	66	1.53
				0.60	G4 v G2	260	65	1.51
				0.60	G4 v G3	264	66	1.53
0.0	0.0	0.0	0.7		Omnibus	120	30	
				0.00	G1 v G2	*	*	
				0.00	G2 v G3	*	*	
				0.00	G3 v G1	*	*	
				0.70	G4 v G1	196	49	1.63
				0.70	G4 v G2	188	47	1.57
				0.70	G4 v G3	192	48	1.60
0.0	0.0	0.0	0.8		Omnibus	96	24	
				0.00	G1 v G2	*	*	
				0.00	G2 v G3	*	*	
				0.00	G3 v G1	*	*	
				0.80	G4 v G1	148	37	1.54
				0.80	G4 v G2	148	37	1.54
				0.80	G4 v G3	144	36	1.50

0.0	0.0	0.2	0.2		Omnibus	1112	278	
				0.00	G1 v G2	*	*	
				0.20	G2 v G3	2252	563	2.03
				0.20	G3 v G1	2316	579	2.08
				0.20	G4 v G1	2336	584	2.10
				0.20	G4 v G2	2348	587	2.11
				0.00	G4 v G3	*	*	
0.0	0.0	0.3	0.3		Omnibus	488	122	
				0.00	G1 v G2	*	*	
				0.30	G2 v G3	1016	254	2.08
				0.30	G3 v G1	1052	263	2.16
				0.30	G4 v G1	1032	258	2.11
				0.30	G4 v G2	1024	256	2.10
				0.00	G4 v G3	*	*	
0.0	0.0	0.4	0.4		Omnibus	280	70	
				0.00	G1 v G2	*	*	
				0.40	G2 v G3	576	144	2.06
				0.40	G3 v G1	588	147	2.10
				0.40	G4 v G1	580	145	2.07
				0.40	G4 v G2	576	144	2.06
				0.00	G4 v G3	*	*	
0.0	0.0	0.5	0.5		Omnibus	180	45	
				0.00	G1 v G2	*	*	
				0.50	G2 v G3	376	94	2.09
				0.50	G3 v G1	376	94	2.09
				0.50	G4 v G1	372	93	2.07
				0.50	G4 v G2	380	95	2.11
				0.00	G4 v G3	*	*	
0.0	0.0	0.6	0.6		Omnibus	128	32	
				0.00	G1 v G2	*	*	
				0.60	G2 v G3	268	67	2.09
				0.60	G3 v G1	264	66	2.06
				0.60	G4 v G1	260	65	2.03
				0.60	G4 v G2	252	63	1.97
				0.00	G4 v G3	*	*	
0.0	0.0	0.7	0.7		Omnibus	92	23	
				0.00	G1 v G2	*	*	
				0.70	G2 v G3	192	48	2.09
				0.70	G3 v G1	188	47	2.04
				0.70	G4 v G1	196	49	2.13
				0.70	G4 v G2	192	48	2.09
				0.00	G4 v G3	*	*	

0.0	0.0	0.8	0.8		Omnibus	72	18	
				0.00	G1 v G2	*	*	
				0.80	G2 v G3	148	37	2.06
				0.80	G3 v G1	148	37	2.06
				0.80	G4 v G1	144	36	2.00
				0.80	G4 v G2	148	37	2.06
				0.00	G4 v G3	*	*	
0.0	0.2	0.2	0.4		Omnibus	556	139	
				0.20	G1 v G2	2312	578	4.16
				0.00	G2 v G3	*	*	
				0.20	G3 v G1	2328	582	4.19
				0.40	G4 v G1	588	147	1.06
				0.20	G4 v G2	2332	583	4.19
				0.20	G4 v G3	2348	587	4.22
0.0	0.3	0.3	0.6		Omnibus	248	62	
				0.30	G1 v G2	1048	262	4.23
				0.00	G2 v G3	*	*	
				0.30	G3 v G1	1044	261	4.21
				0.60	G4 v G1	264	66	1.06
				0.30	G4 v G2	1044	261	4.21
				0.30	G4 v G3	1056	264	4.26
0.0	0.4	0.4	0.8		Omnibus	140	35	
				0.40	G1 v G2	596	149	4.26
				0.00	G2 v G3	*	*	
				0.40	G3 v G1	580	145	4.14
				0.80	G4 v G1	148	37	1.06
				0.40	G4 v G2	584	146	4.17
				0.40	G4 v G3	584	146	4.17
0.0	0.0	0.2	0.4		Omnibus	412	103	
				0.00	G1 v G2	*	*	
				0.20	G2 v G3	2272	568	5.51
				0.20	G3 v G1	2336	584	5.67
				0.40	G4 v G1	592	148	1.44
				0.40	G4 v G2	580	145	1.41
				0.20	G4 v G3	2320	580	5.63
0.0	0.0	0.3	0.6		Omnibus	180	45	
				0.00	G1 v G2	*	*	
				0.30	G2 v G3	1048	262	5.82
				0.30	G3 v G1	1036	259	5.76
				0.60	G4 v G1	264	66	1.47
				0.60	G4 v G2	256	64	1.42
				0.30	G4 v G3	1028	257	5.71

0.0	0.0	0.4	0.8		Omnibus	104	26	
				0.00	G1 v G2	*	*	
				0.40	G2 v G3	580	145	5.58
				0.40	G3 v G1	588	147	5.65
				0.80	G4 v G1	148	37	1.42
				0.80	G4 v G2	144	36	1.38
				0.40	G4 v G3	592	148	5.69
0.0	0.2	0.4	0.6		Omnibus	224	56	
				0.20	G1 v G2	2320	580	10.36
				0.20	G2 v G3	2304	576	10.29
				0.40	G3 v G1	604	151	2.70
				0.60	G4 v G1	260	65	1.16
				0.40	G4 v G2	584	146	2.61
				0.20	G4 v G3	2280	570	10.18
0.0	0.3	0.6	0.9		Omnibus	100	25	
				0.30	G1 v G2	1032	258	10.32
				0.30	G2 v G3	1052	263	10.52
				0.60	G3 v G1	256	64	2.56
				0.90	G4 v G1	120	30	1.20
				0.60	G4 v G2	260	65	2.60
				0.30	G4 v G3	1020	255	10.20
0.0	0.0	0.4	0.6		Omnibus	168	42	
				0.00	G1 v G2	*	*	
				0.40	G2 v G3	588	147	3.50
				0.40	G3 v G1	592	148	3.52
				0.60	G4 v G1	260	65	1.55
				0.60	G4 v G2	260	65	1.55
				0.20	G4 v G3	2300	575	13.69
0.0	0.0	0.6	0.9		Omnibus	76	19	
				0.00	G1 v G2	*	*	
				0.60	G2 v G3	260	65	3.42
				0.60	G3 v G1	268	67	3.53
				0.90	G4 v G1	120	30	1.58
				0.90	G4 v G2	124	31	1.63
				0.30	G4 v G3	1056	264	13.89
0.0	0.2	0.2	0.5		Omnibus	352	88	
				0.20	G1 v G2	2320	580	6.59
				0.00	G2 v G3	*	*	
				0.20	G3 v G1	2332	583	6.63
				0.50	G4 v G1	384	96	1.09
				0.30	G4 v G2	1040	260	2.95
				0.30	G4 v G3	1028	257	2.92

0.0	0.2	0.2	0.6		Omnibus	236	59	
				0.20	G1 v G2	2340	585	9.92
				0.00	G2 v G3	*	*	
				0.20	G3 v G1	2332	583	9.88
				0.60	G4 v G1	276	69	1.17
				0.40	G4 v G2	580	145	2.46
				0.40	G4 v G3	588	147	2.49
0.0	0.2	0.2	0.7		Omnibus	172	43	
				0.20	G1 v G2	2360	590	13.72
				0.00	G2 v G3	*	*	
				0.20	G3 v G1	2352	588	13.67
				0.70	G4 v G1	192	48	1.12
				0.50	G4 v G2	380	95	2.21
				0.50	G4 v G3	376	94	2.19
0.0	0.2	0.2	0.8		Omnibus	128	32	
				0.20	G1 v G2	2304	576	18.00
				0.00	G2 v G3	*	*	
				0.20	G3 v G1	2320	580	18.13
				0.80	G4 v G1	148	37	1.16
				0.60	G4 v G2	256	64	2.00
				0.60	G4 v G3	264	66	2.06
0.0	0.3	0.3	0.7		Omnibus	180	45	
				0.30	G1 v G2	1028	257	5.71
				0.00	G2 v G3	*	*	
				0.30	G3 v G1	1020	255	5.67
				0.70	G4 v G1	196	49	1.09
				0.40	G4 v G2	596	149	3.31
				0.40	G4 v G3	584	146	3.24
0.0	0.3	0.3	0.8		Omnibus	140	35	
				0.30	G1 v G2	1064	266	7.60
				0.00	G2 v G3	*	*	
				0.30	G3 v G1	1048	262	7.49
				0.80	G4 v G1	148	37	1.06
				0.50	G4 v G2	384	96	2.74
				0.50	G4 v G3	380	95	2.71
0.0	0.2	0.4	0.7		Omnibus	172	43	
				0.20	G1 v G2	2260	565	13.14
				0.20	G2 v G3	2364	591	13.74
				0.40	G3 v G1	588	147	3.42
				0.70	G4 v G1	192	48	1.12
				0.50	G4 v G2	364	91	2.12
				0.30	G4 v G3	1024	256	5.95

0.0	0.2	0.5	0.7		Omnibus	156	39	
				0.20	G1 v G2	2292	573	14.69
				0.30	G2 v G3	1072	268	6.87
				0.50	G3 v G1	380	95	2.44
				0.70	G4 v G1	192	48	1.23
				0.50	G4 v G2	376	94	2.41
				0.20	G4 v G3	2320	580	14.87
0.0	0.2	0.4	0.8		Omnibus	128	32	
				0.20	G1 v G2	2272	568	17.75
				0.20	G2 v G3	2352	588	18.38
				0.40	G3 v G1	592	148	4.63
				0.80	G4 v G1	148	37	1.16
				0.60	G4 v G2	256	64	2.00
				0.40	G4 v G3	592	148	4.63
0.0	0.2	0.6	0.8		Omnibus	112	28	
				0.20	G1 v G2	2356	589	21.04
				0.40	G2 v G3	592	148	5.29
				0.60	G3 v G1	260	65	2.32
				0.80	G4 v G1	148	37	1.32
				0.60	G4 v G2	264	66	2.36
				0.20	G4 v G3	2300	575	20.54
0.0	0.3	0.5	0.8		Omnibus	136	34	
				0.30	G1 v G2	1008	252	7.41
				0.20	G2 v G3	2340	585	17.21
				0.50	G3 v G1	372	93	2.74
				0.80	G4 v G1	148	37	1.09
				0.50	G4 v G2	372	93	2.74
				0.30	G4 v G3	1036	259	7.62
0.0	0.2	0.5	0.8		Omnibus	124	31	
				0.20	G1 v G2	2372	593	19.13
				0.30	G2 v G3	1068	267	8.61
				0.50	G3 v G1	384	96	3.10
				0.80	G4 v G1	144	36	1.16
				0.60	G4 v G2	260	65	2.10
				0.30	G4 v G3	1028	257	8.29
0.0	0.0	0.2	0.5		Omnibus	272	68	
				0.00	G1 v G2	*	*	
				0.20	G2 v G3	2328	582	8.56
				0.20	G3 v G1	2348	587	8.63
				0.50	G4 v G1	380	95	1.40
				0.50	G4 v G2	372	93	1.37
				0.30	G4 v G3	1044	261	3.84

0.0	0.0	0.2	0.6		Omnibus	188	47	
				0.00	G1 v G2	*	*	
				0.20	G2 v G3	2312	578	12.30
				0.20	G3 v G1	2312	578	12.30
				0.60	G4 v G1	260	65	1.38
				0.60	G4 v G2	264	66	1.40
				0.40	G4 v G3	580	145	3.09
0.0	0.0	0.2	0.7		Omnibus	136	34	
				0.00	G1 v G2	*	*	
				0.20	G2 v G3	2320	580	17.06
				0.20	G3 v G1	2368	592	17.41
				0.70	G4 v G1	188	47	1.38
				0.70	G4 v G2	188	47	1.38
				0.50	G4 v G3	372	93	2.74
0.0	0.0	0.2	0.8		Omnibus	108	27	
				0.00	G1 v G2	*	*	
				0.20	G2 v G3	2292	573	21.22
				0.20	G3 v G1	2300	575	21.30
				0.80	G4 v G1	152	38	1.41
				0.80	G4 v G2	144	36	1.33
				0.60	G4 v G3	256	64	2.37
0.0	0.0	0.3	0.5		Omnibus	248	62	
				0.00	G1 v G2	*	*	
				0.30	G2 v G3	1032	258	4.16
				0.30	G3 v G1	1036	259	4.18
				0.50	G4 v G1	384	96	1.55
				0.50	G4 v G2	376	94	1.52
				0.20	G4 v G3	2324	581	9.37
0.0	0.0	0.5	0.7		Omnibus	116	29	
				0.00	G1 v G2	*	*	
				0.50	G2 v G3	384	96	3.31
				0.50	G3 v G1	372	93	3.21
				0.70	G4 v G1	192	48	1.66
				0.70	G4 v G2	196	49	1.69
				0.20	G4 v G3	2268	567	19.55
0.0	0.0	0.6	0.8		Omnibus	88	22	
				0.00	G1 v G2	*	*	
				0.60	G2 v G3	260	65	2.95
				0.60	G3 v G1	268	67	3.05
				0.80	G4 v G1	148	37	1.68
				0.80	G4 v G2	148	37	1.68
				0.20	G4 v G3	2316	579	26.32

0.0	0.0	0.3	0.7		Omnibus	136	34	
				0.00	G1 v G2	*	*	
				0.30	G2 v G3	1052	263	7.74
				0.30	G3 v G1	1044	261	7.68
				0.70	G4 v G1	196	49	1.44
				0.70	G4 v G2	192	48	1.41
				0.40	G4 v G3	580	145	4.26
0.0	0.0	0.3	0.8		Omnibus	108	27	
				0.00	G1 v G2	*	*	
				0.30	G2 v G3	1032	258	9.56
				0.30	G3 v G1	1064	266	9.85
				0.80	G4 v G1	152	38	1.41
				0.80	G4 v G2	148	37	1.37
				0.50	G4 v G3	368	92	3.41
0.0	0.0	0.4	0.7		Omnibus	128	32	
				0.00	G1 v G2	*	*	
				0.40	G2 v G3	560	140	4.38
				0.40	G3 v G1	596	149	4.66
				0.70	G4 v G1	192	48	1.50
				0.70	G4 v G2	188	47	1.47
				0.30	G4 v G3	1028	257	8.03
0.0	0.0	0.5	0.8		Omnibus	100	25	
				0.00	G1 v G2	*	*	
				0.50	G2 v G3	376	94	3.76
				0.50	G3 v G1	380	95	3.80
				0.80	G4 v G1	148	37	1.48
				0.80	G4 v G2	144	36	1.44
				0.30	G4 v G3	1036	259	10.36

Notes. :* Null Hypothesis was true for the given comparison, so no sample size analysis was performed
[a] Relative efficiency is calculated as the Total sample size for the particular comparison divided by the Total Sample Size for the Omnibus Test for the condition.

Pattern 1: Three equal mean differences (and three nil differences)

In situations where three groups had the same standardized mean and a fourth group mean differed, which resulted in three equal mean differences in multiple comparison, the non-null multiple comparisons required larger sample sizes than the omnibus ANOVA. For example, the condition where the pattern of standardized means was 0.0, 0.0, 0.0 and 0.2 (therefore a pattern of mean differences of 0.0, 0.0, 0.0, 0.2, 0.2 and 0.2) resulted in per group sample sizes of roughly 580 cases to achieve power of .80 for the three multiple comparisons with a standardized mean difference of 0.2, which was compared to the 358 cases per group needed to achieve statistical power of .80 for the omnibus test. In fact, the first seven cases listed in Table 1 indicated that no matter how the magnitude of the mean difference, the relative efficiency was approximately 1.58. In other words, in all cases where three groups had the same mean while a fourth group differed, the multiple comparisons required approximately 1.58 times more cases than the omnibus test did in order to achieve power of .80.

Pattern 2: Four equal mean differences (and two nil differences)

In conditions where two groups had the same standardized means while the other two were same, which resulted in four equal mean differences in multiple comparison procedure, the non-null multiple comparisons still required larger sample sizes than the omnibus ANOVA. For example, the condition where the pattern of standardized means was 0.0, 0.0, 0.2 and 0.2 (therefore a pattern of mean differences in multiple comparison of 0.0, 0.0, 0.2, 0.2, 0.2 and 0.2) resulted in per group sample sizes of roughly 580 cases to achieve power of .80 for the four multiple comparisons with a standardized mean difference of 0.2, which was compared to the 278 cases per group needed to achieve statistical power of .80 for the omnibus test. As for the relative efficiency, it was about 2.08. As another example, in the case where the pattern of means was 0.0, 0.0, 0.3 and 0.3 (therefore a pattern of mean differences in multiple comparison of 0.0, 0.0, 0.3, 0.3, 0.3 and 0.3) resulted in per group sample sizes of roughly 260 cases to achieve power of .80 for the four multiple comparisons with a standardized mean difference of 0.3, which was compared to the 122 cases per group needed to achieve statistical power of .80 for the omnibus test. Also, the similar relative efficiency 2.06 was obtained in the pattern of means 0.0, 0.0, 0.8 and 0.8, where 37 cases were required per group in multiple comparison while only 18 cases needed in the omnibus test to meet the power of .80. Altogether, stated in another way, in all cases where four equal mean differences were shown in multiple comparison procedure, roughly 2.08 times more cases required than the omnibus test did in order to achieve power of .80.

Pattern 3: Four equal mean differences, with a fifth twice as large (and one nil)

In conditions where there were four equal mean differences in multiple comparison while a fifth group mean was twice as large, the four equal mean differences required a much larger sample size than the overall test, while the fifth group mean difference, which was twice as large, required roughly same as the omnibus test. For example, in the case where the pattern of means was 0.0, 0.2, 0.2 and 0.4 (therefore a pattern of mean differences in multiple comparison of 0.0, 0.2, 0.2, 0.2, 0.2 and 0.4) resulted in per group sample sizes of roughly 139 cases to achieve power of .80 in the omnibus test. While in multiple comparison procedure, about 580 cases required for the mean difference 0.2 per group, 147 cases for the mean difference 0.4, the relative efficiency would be 4.20 and 1.06 respectively. Also, if we look at the pattern of means of 0.0, 0.3, 0.3 and 0.6, the same pattern of relative efficiency would be found, so does the pattern of means of 0.0, 0.4, 0.4 and 0.8. In fact, if the pattern of means was not centered or standardized, for example, a pattern of group means of 0.5, 0.7, 0.7 and 0.9 (therefore a pattern of mean differences of 0.0, 0.2, 0.2, 0.2, 0.2 and 0.4) would end up with the similar relative efficiency as the pattern of 0.0, 0.2, 0.2, and 0.4. Altogether, for cases of four equal mean differences and a fifth twice as large in multiple comparison procedure, the four equal mean differences would require about 4.20 times more cases than the omnibus test did in order to achieve power of 0.8, while for the mean difference, which was twice as large, the relative efficiency would be 1.06.

Pattern 4: Three equal mean differences and two twice as large (and one nil)

For example, in the case where the pattern of means was 0.0, 0.0, 0.2 and 0.4 (therefore a pattern of mean differences in multiple comparison of 0.0, 0.2, 0.2, 0.2, 0.4 and 0.4) resulted in per group sample sizes of roughly 103 cases to achieve power of .80 in the omnibus test. As for the mean difference 0.2, it required roughly 580 cases which resulted in the relative efficiency about 5.60, while for the mean difference 0.4, it required approximately 148 per group and the relative efficiency was 1.44. Overall, the pattern of relative efficiency would be consistent if we look at the other two similar patterns of means (0.0, 0.0, 0.3 and 0.6) and (0.0, 0.0, 0.4 and 0.8). The relative efficiency, as compared to the omnibus test, was about 5.60 for the three equal mean differences and 1.44 for the other two means which are twice as large.

Pattern 5: Three equal mean differences, two twice as large and one 3 times larger

For example, a pattern of group means of 0.0, 0.2, 0.4 and 0.6 (therefore a pattern of mean differences in multiple comparison of 0.2, 0.2, 0.2, 0.4, 0.4 and 0.6) resulted in per group sample sizes of roughly 56 cases to achieve power of .80 in the omnibus test. As for the three equal mean difference of 0.2, it required roughly 580 cases which resulted in the relative efficiency about 10.28, while for the mean difference of 0.4, it required approximately 148 per group and the relative efficiency was 2.65, as for the mean difference 0.6, which was three times as large, 65 cases required per group and the relative efficiency was about 1.16. Take the other pattern of 0.0, 0.3, 0.6 and 0.9 as another example (which was run as an additional supplemental analysis for confirmation of this result). Obviously, there were three equal mean differences, two twice as large and one three times. As a result, the relative efficiency was consistent, which was about 10.30 for the three equal mean differences of 0.3, 2.58 for the two twice as large of 0.6 and 1.20 for the mean difference of 0.9 which was three times as large. Also, the pattern such as 0.2, 0.4, 0.6, and 0.8 would end up with similar relative efficiency as the patterns in our examples.

Pattern 6: One smaller mean difference, two twice as large, two 3 times larger (and one nil)

For example, a pattern of group means of 0.0, 0.0, 0.4 and 0.6 (therefore a pattern of mean differences in multiple comparison of 0.0, 0.2, 0.4, 0.4, 0.6 and 0.6) resulted in per group sample sizes of roughly 42 cases to achieve power of .80 in the omnibus test. As for the small mean difference of 0.2, it required roughly 580 cases which resulted in the relative efficiency about 13.70, while for the mean difference of 0.4 which was twice as large, it required approximately 148 per group and the relative efficiency was 3.50, as for the mean difference 0.6, which was three times as large, 65 cases required per group and the relative efficiency was about 1.55. Take the other pattern of 0.0, 0.0, 0.6 and 0.9 as another example, which had a similar pattern of mean differences as the pattern of 0.0, 0.0, 0.4, and 0.6 (run as a supplemental analysis to confirm this result more generally). As a result, the relative efficiency was consistent, which was about 13.89 for one small equal mean difference of 0.3, 3.48 for the mean difference of 0.6 and 1.60 for the mean difference of 0.9 which was three times as large.

Absolute Mean Difference Effect Sizes

Ultimately, there were relatively consistent required sample sizes for absolute group mean differences regardless of the pattern of means. That is, no matter what the pattern of means across the four groups (above), the same sample size was required for any given absolute mean difference (see Table 2). For example, when the sample size required for the mean difference effect size 0.2 in the multiple comparison procedure, it will always be roughly 2332 total cases, that is, about 580 cases per group needed to achieve the statistical power of .80 no matter what kind of pattern of the four group means. As for the mean difference effect size 0.3, totally 1040 cases are required. However, note that the relationship between the sample size for the ANOVA omnibus F test and the multiple comparisons does depend on the patterns, because the patterns play a role in the calculation of Cohen's *f* effect size for ANOVA.

Conclusions

These results clearly show that adequate statistical power for the omnibus ANOVA *F* test does not guarantee adequate statistical power for given pairwise MCPs performed post hoc. Therefore, when a researcher is considering sample size, it may not be sufficient to set sample size for the omnibus test being performed. More important, researchers should consider the post hoc multiple comparison procedure. In this context, the sample size for the absolute mean difference effect size (Table 2) was the most important finding in this paper which could provide researchers with some hints to decide the sufficient sample size not only for the omnibus *F* test but for the post hoc MCPs.

Recommendations

Based on the results presented, there are certain specific recommendations that can be made concerning sample sizes researchers should use in ANOVA with four groups. It must be remembered that these results were limited to Tukey HSD comparisons performed using statistical power of .80. However, we would expect that other post hoc comparisons would show similar patterns of results—just with different sample sizes.

A Heuristic Example

A researcher may expect a pattern of means across four groups of 0.3, 0.3, 0.5, and 0.7. How is the sample size decided in order to achieve significant results not only in the omnibus *F* test but in Tukey HSD multiple comparison? There are two ways referring to Table 1 and Table 2. First, the pattern of means of 0.3, 0.3, 0.5, and 0.7 could be centered into 0.0, 0.0, 0.2, and 0.4. According to Table 1, about 412 cases (103 cases per group) are required for the omnibus *F* test, while about 2320 cases are required (580 cases per group) for the standardized mean difference of 0.2, and 590 cases (148 cases per group) are needed for the mean difference of 0.4 in the multiple comparison procedure. Hence, in this case, 580 cases per group should be used, which will provide significant results in the MCPs at the statistical power desired if all tests are important.

Alternatively, the pairwise mean difference of the pattern of 0.3, 0.3, 0.5, and 0.7, was 0.0, 0.2, 0.2, 0.2, 0.2, and 0.4. Based on Table 2, 580 cases per group are required for the absolute mean difference of 0.2 and 148 cases per group are needed for the mean difference of 0.4. Since the required sample size for the MCP was larger than that of omnibus *F* test, 580 cases per group or totally about 2320 cases are required

Table 2. Sample size required for statistical power of .80 for the Tukey HSD Multiple Comparison Procedure given specific standardized mean differences, no matter what the pattern of group means.

Standardized Mean Difference Effect Size	Total Sample Size	Per Group Sample Size
0.2	2320	580
0.3	1040	260
0.4	590	148
0.5	375	94
0.6	262	66
0.7	190	48
0.8	146	37
0.9	120	30

in this design so as to achieve significant results with the power of .80. However, the researcher may determine that only certain of the multiple comparisons are most important for the research questions and choose sample sizes to provide power for those comparisons of most interest. That is, in this example, perhaps the researcher is much more interested in the differences between Group 4 (expected 0.7 mean) and Groups 1 and 2 (expected 0.3 means), for an expected standardized mean difference effect size of 0.4. Because the researcher is less interested in the standardized group differences of 0.2, sample size can be chosen for the 0.4 comparisons of most interest.

Finally, the researcher might use the relative efficiency column in Table 1 to estimate necessary sample sizes for the multiple comparisons based on the omnibus *F* test. That is, the 0.2 standardized mean differences in this expected population condition require roughly 5.6 times more cases that the omnibus *F* test. However, the 0.4 standardized mean differences require roughly 1.4 times more cases than the omnibus test. Unfortunately, these patterns are dependent on the patterns of the means across groups and cannot easily be generalized to all situations. But if the researcher knows the expected pattern of means (and therefore pattern of standardized mean differences), even without knowing the actual means, this approach can be useful. This example falls into Pattern 3 described above.

Discussion

Similar relative efficiency results were obtained at different power levels (0.7 and 0.9) using the MC4G program. The group mean of 0.9 was also taken into consideration in pattern 5 and pattern 6, which confirmed the consistency of relative efficiency. Therefore, the authors recommend that for other power levels, the Relative Efficiency approach described above should be used. This paper provides several important implications for researchers. Most importantly, the appropriate sample sizes for MCPs will help investigators be able to obtain results from pairwise group comparisons for which they are comfortable that statistical power was adequate. Also, because most scholars have only focused on sample size selection that ensures enough power in the omnibus ANOVA *F* test, this study will fill an important gap regarding sample size determination in MCPs. Finally, the MC4G Monte Carlo program is available to researchers for to use for power and sample size analyses as well as other related simulations. Monte Carlo methods may become especially useful in helping to determine sample sizes in conditions where exact analytical methods will not work. Certainly, the MC4G program used here has limitations, but similar Monte Carlo simulations could be performed in languages such as R or Python with fewer such limitations.

Admittedly, the results in this paper were only based on Tukey HSD in MCPs, other procedures designed, such as Bonferroni, Games-Howell, Scheffé, and Dunnett tests, might be included in future research. We have reason to believe that other MCPs will show similar patterns of differences from the omnibus tests, but the sample size details are likely to differ because of well-known differences in power among MCPs. In addition, Brooks and Johanson (2011) determined the sample size for MCPs in ANOVA across three groups, and this paper only includes four groups, so additional groups should be considered in future research to see whether the patterns of the sample sizes are consistent a broader range of mean differences.

References

Barnette, J. J., & McLean, J. E. (1999, April). *Choosing a multiple comparison procedure based on alpha* (ED430047). ERIC. https://eric.ed.gov/?id=ED430047

Brewer, J. K., & Sindelar, P. T. (1988). Adequate sample size: A priori and post hoc considerations. *The Journal of Special Education*, *21*(4), 74-84. DOI:10.1177/002246698802100409

Brooks, G. P. (2018). *MC4G: Monte Carlo Analyses for up to 4 Groups*. [Computer software and manuals]. Retrieved from https://people.ohio.edu/brooksg/#MC4G

Brooks, G. P., & Johanson, G. A. (2011). Sample size considerations for multiple comparison procedures in ANOVA. *Journal of Modern Applied Statistical Methods*, *10*(1), 10. DOI:10.22237/jmasm/1304222940

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Klockars, A. J. & Hancock, G. R. (1998). A more powerful post hoc multiple comparison procedure in analysis of variance. *Journal of Educational and Behavioral Statistics*, *23*(3), 279-289. DOI:10.2307/1165249

Levin, J. R. (1975). Determining sample size for planned and post hoc analysis of variance comparisons. *Journal of Educational Measurement*, *12*, 99-108. https://www.jstor.org/stable/1434034

Light, R. J., Singer, J. D., & Willett, J. B. (1990). *By design: Planning research on higher education*. Cambridge, MA: Harvard University.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. Proceedings of the National Academy of Sciences, 115(11), 2600–2606. DOI:10.1073/pnas.1708274114

Olejnik, S. F. (1984). Planning educational research: Determining the necessary sample size. *The Journal of Experimental Education*, *53*, 40-48. https://www.jstor.org/stable/20151569

Rubin, M., & Donkin, C. (2022). Exploratory hypothesis tests can be more compelling than confirmatory hypothesis tests. *Philosophical Psychology*, 1-29. DOI:10.1080/09515089.2022.2113771

Raffle, H., & Brooks, G. P. (2005). Using Monte Carlo software to teach abstract statistical concepts: A case study. *Teaching of Psychology, 32*(3), 193-195. DOI:10.1207/s15328023top3203_12

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. Perspectives on Psychological Science, 7(6), 632–638. DOI:10.1177/1745691612463078

Wilcox, R. R. (2002). Understanding the practical advantages of modern ANOVA methods. *Journal of Clinical Child and Adolescent Psychology*, *31*(3), 399-412. DOI:10.1207/153744202760082667

Send correspondence to:

Gordon Brooks
Ohio University
Email: brooksg@ohio.edu