Sample Size Determination for (Planned) Post Hoc Multiple Comparisons in One-Way ANOVA

Frank Oppong

Yuqing Liu

Ohio University

Nina Adjanin

Northwest Missouri State University

George A. Johanson

Gordon P. Brooks

Ohio University

Roundtable Paper presented at the 2023 Annual Meeting of the American Educational Research Association, Philadelphia, PA.

Sample Size Determination for (Planned) Post Hoc Multiple Comparisons in One-Way ANOVA

Abstract

Most approaches to sample size determination in one-way ANOVA are based on expected variation in means, but these approaches provide power only for the omnibus *F* test. Unfortunately, sufficient sample size for the omnibus ANOVA test does not guarantee adequate statistical power for post hoc multiple comparisons that most researchers plan to perform following a significant ANOVA (that is, not many stop after reporting a significant omnibus test). This Monte Carlo study investigated the sample sizes needed for the smallest, most-interesting post hoc mean comparison expected to be performed following a significant ANOVA. We identified sample size rules that can be applied generally for Tukey-Kramer and Games-Howell MCPs across many numbers of groups and effect size of most interest conditions.

Objectives

As they plan their studies, researchers often attempt to identify the minimum number of participants or cases needed to test a hypothesis of interest. In ANOVA, the choice of sample size is impacted by the strategy that researchers choose (e.g., omnibus test power, any-pair power, all-pair power). Brooks and Johanson (2011) found that when one-way 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. Therefore, researchers should ensure sufficient sample sizes for the most salient statistical analyses they plan to perform—what we might call “specific-pair” power. Following from this perspective, the purpose of this paper is to demonstrate a rule that can be used to determine sample sizes for two of the most common post hoc multiple comparison procedures (MCP), Tukey-Kramer (hereafter called Tukey as is common in most programs) and Games-Howell. Pairwise multiple comparisons among 3 to 9 groups across several effect sizes were studied in order to develop and confirm this approach.

Perspectives

Barnette and McLean (1999) wrote that many scholars conduct post hoc pairwise MCPs after significant omnibus one-way ANOVA *F* tests, but this results in concerns over the control of Type I error for multiple hypothesis tests. MCPs are used to control the family-wise Type I error (Klockars & Hancock, 1998). Tukey HSD is highly recommended when variances are homogeneous and Games-Howell when homogeneity of variances is violated. Both can be used as unprotected tests (Barnette & McLean, 1999) and therefore do not need to be protected by a statistically significant ANOVA (i.e., can be used without first testing the omnibus ANOVA).

Most studies that use one-way ANOVA use the approach based on Cohen (1988), in which effect size is based on the variation in means and sample size is calculated only for the omnibus test. Others (e.g., Hsu, 1999, Levin, 1975) have proposed other approaches, but the omnibus ANOVA power analysis approach remains most common. However, Brooks and Johanson (2011) demonstrated that sample sizes that provided adequate statistical power for three-group omnibus ANOVA *F* tests are not sufficient for the pairwise comparisons typically performed post hoc. They also showed that the patterns of means do not impact the sample size determination when seeking the sample size required for the specific-pair comparison-of-most-interest. That is, researchers should estimate the two-group effect size (i.e., Cohen's *d*) for the most important pairwise comparison or for the pairwise comparison with the smallest expected Cohen's *d*.

Studies will sometimes result in a statistically significant omnibus test but no statistically significant MCPs. This will occur at times because the omnibus test reflects that a non-pairwise comparison is significant. However, it will sometimes occur because there is simply not enough power for the adjusted-alpha MCP being performed by the researcher. In this case, the specific-pair comparison-of-most-interest approach to sample size selection is useful. There have been few studies that focused on the sample size determination for MCPs. Therefore, this paper will focus on the sample size selection for researchers planning to perform a post hoc Tukey or Games-Howell MCP after a statistically significant ANOVA.

Methods

Monte Carlo programs were written in R to perform the computer simulations required for two phases of the study.

Phase 1 Methods and Data Sources

In Phase 1, a program used Monte Carlo techniques to find sample sizes that meet given statistical power levels for three to six groups. Sample sizes were increased until power reached the given desired values. Power analyses were performed using 2,000 replications due to the many repeated analyses required as sample sizes were changed using built-in R functions. At lower levels of power multiple cases were added to the analyses, but as the power level increased past 50%, only one case was added to the datasets until the desired power levels were reached. For example, when power reached 70%, that sample size was recorded and cases continued to be added until power reached 80% and then 90%.

Following the Monte Carlo simulations, the results were examined for patterns that could be useful for a sample size rule that would be generally acceptable across many conditions without the need for tables. For example, the sample sizes for number of groups were compared (e.g., the ratio of sample sizes for four groups relative to three). Sample sizes obtained for given power levels were compared to the required sample sizes for independent t tests. Additionally, sample sizes were compared to those required for independent t tests when a Bonferroni adjustment was made to alpha based on the total number of possible comparisons (e.g., with five groups, there are ten possible comparisons).

Phase 1 Results

In total, 96 combinations of number of groups (3 to 6), Cohen's d effect sizes (0.2 to 0.9), and power levels (.70 to .90) were used. Table 1 shows the average ratio of the sample sizes obtained from Monte Carlo simulations for the Tukey, Games-Howell, and Scheffe MCPs to the known sample sizes required for the independent t test at alpha of .05 and the independent t test at the appropriate Bonferroni-adjusted alpha.

Across all 96 conditions, the ratio of sample sizes required by Games-Howell to the Bonferroni-adjusted t test was .944. In other words, the sample size required for Games-Howell was 94.4% as large as that required by the independent t test with Bonferroni adjustment. Both Tukey and Games-Howell have consistent relationships with the sample sizes needed for the Bonferroni-adjusted t tests. The range of ratios for Tukey was 0.89 to 0.97 while the range for Games-Howell was 0.91 to 0.99. However, sample sizes for Games-Howell were more symmetric around its ratio than were the sample sizes for Tukey (see Figure 1). Games-Howell had 74 conditions within ± 0.015 of the Bonferroni sample size (with 11 above and 11 below), while Tukey had 70 (with 14 below and 12 above).

```
> VERTSORT7
```

	Power	g	m	f	n_F	n_t	n_b	avgG	G_B	avgT	T_B	avgS	S_B
1	0.7	3	0.9	0.424	16	17	23	22.00000	0.9565217	22.00000	0.9565217	23.00000	1.0000000
2	0.7	3	0.8	0.377	20	21	29	26.50000	0.9137931	26.50000	0.9137931	28.50000	0.9827586
3	0.7	3	0.7	0.330	25	27	37	34.50000	0.9324324	34.00000	0.9189189	37.50000	1.0135135
4	0.7	3	0.6	0.283	34	36	49	45.00000	0.9183673	45.50000	0.9285714	50.00000	1.0204082
5	0.7	3	0.5	0.236	48	51	70	67.50000	0.9642857	66.00000	0.9428571	70.00000	1.0000000
6	0.7	3	0.4	0.189	74	79	108	103.00000	0.9537037	102.00000	0.9444444	109.00000	1.0092593
7	0.7	3	0.3	0.141	130	139	191	178.50000	0.9345550	180.00000	0.9424084	197.00000	1.0314136
8	0.7	3	0.2	0.094	290	310	428	388.00000	0.9065421	391.50000	0.9147196	431.00000	1.0070093
25	0.7	4	0.9	0.390	16	17	27	25.00000	0.9259259	24.00000	0.8888889	29.00000	1.0740741
26	0.7	4	0.8	0.346	20	21	34	32.00000	0.9411765	31.00000	0.9117647	35.66667	1.0490196
27	0.7	4	0.7	0.303	25	27	43	40.33333	0.9379845	39.66667	0.9224806	46.33333	1.0775194
28	0.7	4	0.6	0.260	34	36	58	54.00000	0.9310345	54.00000	0.9310345	63.33333	1.0919540
29	0.7	4	0.5	0.217	48	51	82	77.66667	0.9471545	76.66667	0.9349593	88.00000	1.0731707
30	0.7	4	0.4	0.173	75	79	127	117.66667	0.9265092	116.66667	0.9186352	136.33333	1.0734908
31	0.7	4	0.3	0.130	132	139	225	208.00000	0.9244444	208.00000	0.9244444	243.66667	1.0829630
32	0.7	4	0.2	0.087	295	310	502	469.00000	0.9342629	463.66667	0.9236388	539.00000	1.0737052
49	0.7	5	0.9	0.360	16	17	30	28.00000	0.9333333	27.00000	0.9000000	33.50000	1.1166667
50	0.7	5	0.8	0.320	20	21	37	35.50000	0.9594595	34.00000	0.9189189	41.25000	1.1148649
51	0.7	5	0.7	0.280	26	27	48	44.75000	0.9322917	44.50000	0.9270833	54.50000	1.1354167
52	0.7	5	0.6	0.240	35	36	64	59.50000	0.9296875	58.00000	0.9062500	73.00000	1.1406250
53	0.7	5	0.5	0.200	50	51	91	87.00000	0.9560440	83.25000	0.9148352	104.50000	1.1483516
54	0.7	5	0.4	0.160	77	79	141	133.25000	0.9450355	132.00000	0.9361702	160.75000	1.1400709
55	0.7	5	0.3	0.120	136	139	249	233.75000	0.9387550	233.75000	0.9387550	284.00000	1.1405622
56	0.7	5	0.2	0.080	304	310	557	518.50000	0.9308797	513.25000	0.9214542	642.00000	1.1526032
73	0.7	6	0.9	0.335	17	17	32	30.40000	0.9500000	29.00000	0.9062500	38.00000	1.1875000
74	0.7	6	0.8	0.298	21	21	40	37.60000	0.9400000	36.60000	0.9150000	47.20000	1.1800000
75	0.7	6	0.7	0.261	27	27	52	47.80000	0.9192308	47.60000	0.9153846	61.80000	1.1884615
76	0.7	6	0.6	0.224	36	36	69	66.20000	0.9594203	64.80000	0.9391304	84.20000	1.2202899
77	0.7	6	0.5	0.186	52	51	98	93.40000	0.9530612	92.80000	0.9469388	119.00000	1.2142857
78	0.7	6	0.4	0.149	80	79	152	141.60000	0.9315789	140.60000	0.9250000	185.40000	1.2197368
79	0.7	6	0.3	0.112	141	139	269	245.80000	0.9137546	242.40000	0.9011152	328.40000	1.2208178
80	0.7	6	0.2	0.075	315	310	601	547.40000	0.9108153	542.80000	0.9031614	725.20000	1.2066556

```
> VERTSORT8
```

	Power	g	m	f	n_F	n_t	n_b	avgG	G_B	avgT	T_B	avgS	S_B
9	0.8	3	0.9	0.424	19	21	28	26.50000	0.9464286	26.00000	0.9285714	28.00000	1.000000
10	0.8	3	0.8	0.377	24	26	35	33.50000	0.9571429	33.50000	0.9571429	36.00000	1.028571
11	0.8	3	0.7	0.330	31	34	45	43.00000	0.9555556	42.50000	0.9444444	45.00000	1.000000
12	0.8	3	0.6	0.283	42	45	60	57.50000	0.9583333	57.00000	0.9500000	62.00000	1.033333
13	0.8	3	0.5	0.236	59	64	86	82.50000	0.9593023	82.50000	0.9593023	86.50000	1.005814
14	0.8	3	0.4	0.189	92	100	133	128.50000	0.9661654	128.50000	0.9661654	137.00000	1.030075
15	0.8	3	0.3	0.141	162	176	235	223.50000	0.9510638	222.00000	0.9446809	239.50000	1.019149
16	0.8	3	0.2	0.094	363	394	525	494.00000	0.9409524	498.50000	0.9495238	526.00000	1.001905
33	0.8	4	0.9	0.390	19	21	32	31.00000	0.9687500	30.33333	0.9479167	34.66667	1.083333
34	0.8	4	0.8	0.346	24	26	40	38.66667	0.9666667	37.66667	0.9416667	43.00000	1.075000
35	0.8	4	0.7	0.303	31	34	52	50.00000	0.9615385	48.33333	0.9294872	54.66667	1.051282
36	0.8	4	0.6	0.260	42	45	70	64.66667	0.9238095	64.66667	0.9238095	73.33333	1.047619
37	0.8	4	0.5	0.217	60	64	99	94.66667	0.9562290	93.66667	0.9461279	107.66667	1.087542
38	0.8	4	0.4	0.173	92	100	154	144.66667	0.9393939	144.00000	0.9350649	162.33333	1.054113
39	0.8	4	0.3	0.130	163	176	271	256.00000	0.9446494	256.00000	0.9446494	290.66667	1.072571
40	0.8	4	0.2	0.087	365	394	608	567.00000	0.9325658	567.00000	0.9325658	643.66667	1.058662
57	0.8	5	0.9	0.360	20	21	35	33.50000	0.9571429	32.75000	0.9357143	38.75000	1.107143
58	0.8	5	0.8	0.320	25	26	44	41.50000	0.9431818	41.00000	0.9318182	49.25000	1.119318
59	0.8	5	0.7	0.280	32	34	57	53.75000	0.9429825	53.25000	0.9342105	63.75000	1.118421
60	0.8	5	0.6	0.240	43	45	76	72.00000	0.9473684	71.25000	0.9375000	86.00000	1.131579
61	0.8	5	0.5	0.200	61	64	109	103.25000	0.9472477	101.75000	0.9334862	123.25000	1.130734
62	0.8	5	0.4	0.160	95	100	169	159.25000	0.9423077	159.25000	0.9423077	190.50000	1.127219
63	0.8	5	0.3	0.120	167	176	298	274.25000	0.9203020	274.25000	0.9203020	336.00000	1.127517
64	0.8	5	0.2	0.080	374	394	668	628.75000	0.9412425	627.25000	0.9389970	755.25000	1.130614
81	0.8	6	0.9	0.335	20	21	38	36.60000	0.9631579	35.40000	0.9315789	43.80000	1.152632
82	0.8	6	0.8	0.298	25	26	47	43.80000	0.9319149	43.20000	0.9191489	55.20000	1.174468
83	0.8	6	0.7	0.261	33	34	61	58.00000	0.9508197	56.00000	0.9180328	71.40000	1.170492
84	0.8	6	0.6	0.224	44	45	82	77.60000	0.9463415	76.60000	0.9341463	96.40000	1.175610
85	0.8	6	0.5	0.186	63	64	117	110.60000	0.9452991	109.00000	0.9316239	138.80000	1.186325
86	0.8	6	0.4	0.149	98	100	181	170.60000	0.9425414	169.40000	0.9359116	215.60000	1.191160
87	0.8	6	0.3	0.112	172	176	320	300.00000	0.9375000	298.60000	0.9331250	379.60000	1.186250
88	0.8	6	0.2	0.075	386	394	716	670.60000	0.9365922	669.40000	0.9349162	851.40000	1.189106

> VERTSORT9

	Power	g	m	f	n_F	n_t	n_b	avgG	G_B	avgT	T_B	avgS	S_B
17	0.9	3	0.9	0.424	25	27	35	34.50000	0.9857143	34.00000	0.9714286	36.50000	1.0428571
18	0.9	3	0.8	0.377	31	34	44	43.00000	0.9772727	42.50000	0.9659091	45.50000	1.0340909
19	0.9	3	0.7	0.330	40	44	57	54.50000	0.9561404	53.50000	0.9385965	57.50000	1.0087719
20	0.9	3	0.6	0.283	54	60	77	73.50000	0.9545455	73.50000	0.9545455	78.00000	1.0129870
21	0.9	3	0.5	0.236	77	86	110	107.00000	0.9727273	106.50000	0.9681818	109.00000	0.9909091
22	0.9	3	0.4	0.189	120	133	171	162.50000	0.9502924	162.50000	0.9502924	171.00000	1.0000000
23	0.9	3	0.3	0.141	212	235	302	285.50000	0.9453642	283.00000	0.9370861	305.50000	1.0115894
24	0.9	3	0.2	0.094	476	527	677	630.00000	0.9305761	630.00000	0.9305761	673.50000	0.9948301
41	0.9	4	0.9	0.390	25	27	40	39.00000	0.9750000	38.00000	0.9500000	42.66667	1.0666667
42	0.9	4	0.8	0.346	31	34	50	47.33333	0.9466667	47.00000	0.9400000	53.66667	1.0733333
43	0.9	4	0.7	0.303	40	44	65	61.00000	0.9384615	61.00000	0.9384615	70.00000	1.0769231
44	0.9	4	0.6	0.260	54	60	88	82.33333	0.9356061	82.33333	0.9356061	92.00000	1.0454545
45	0.9	4	0.5	0.217	77	86	125	118.33333	0.9466667	117.33333	0.9386667	132.66667	1.0613333
46	0.9	4	0.4	0.173	120	133	194	181.66667	0.9364261	182.33333	0.9398625	206.66667	1.0652921
47	0.9	4	0.3	0.130	211	235	344	321.66667	0.9350775	318.66667	0.9263566	357.33333	1.0387597
48	0.9	4	0.2	0.087	474	527	770	730.00000	0.9480519	728.33333	0.9458874	808.33333	1.0497835
65	0.9	5	0.9	0.360	25	27	44	42.50000	0.9659091	40.50000	0.9204545	48.00000	1.0909091
66	0.9	5	0.8	0.320	32	34	55	52.25000	0.9500000	51.00000	0.9272727	60.75000	1.1045455
67	0.9	5	0.7	0.280	41	44	71	67.50000	0.9507042	66.50000	0.9366197	78.00000	1.0985915
68	0.9	5	0.6	0.240	55	60	95	89.50000	0.9421053	88.50000	0.9315789	105.25000	1.1078947
69	0.9	5	0.5	0.200	78	86	136	130.25000	0.9577206	128.00000	0.9411765	151.75000	1.1158088
70	0.9	5	0.4	0.160	122	133	211	200.25000	0.9490521	199.50000	0.9454976	236.00000	1.1184834
71	0.9	5	0.3	0.120	215	235	374	353.00000	0.9438503	352.25000	0.9418449	413.25000	1.1049465
72	0.9	5	0.2	0.080	483	527	838	783.75000	0.9352625	783.75000	0.9352625	930.50000	1.1103819
89	0.9	6	0.9	0.335	26	27	47	44.40000	0.9446809	43.20000	0.9191489	52.80000	1.1234043
90	0.9	6	0.8	0.298	32	34	58	55.40000	0.9551724	53.20000	0.9172414	68.00000	1.1724138
91	0.9	6	0.7	0.261	42	44	75	71.60000	0.9546667	70.60000	0.9413333	88.20000	1.1760000
92	0.9	6	0.6	0.224	56	60	101	95.00000	0.9405941	95.00000	0.9405941	117.40000	1.1623762
93	0.9	6	0.5	0.186	80	86	145	138.40000	0.9544828	136.60000	0.9420690	170.00000	1.1724138
94	0.9	6	0.4	0.149	125	133	225	212.00000	0.9422222	209.60000	0.9315556	261.40000	1.1617778
95	0.9	6	0.3	0.112	221	235	398	374.20000	0.9402010	374.40000	0.9407035	468.60000	1.1773869
96	0.9	6	0.2	0.075	496	527	892	833.20000	0.9340807	829.80000	0.9302691	1038.60000	1.1643498

Table 1. Ratio of MCP sample sizes to the t and Bonferroni-adjusted t sample sizes

	TukeyKramer:t-Test	TukeyKramer:Bonf	GamesHowell:t-Test	GamesHowell:Bonf	Scheffe:t-Test	Scheffe:Bonf
Median	1.496	0.935	1.504	0.945	1.756	1.091
Mean	1.485	0.934	1.503	0.944	1.757	1.096
StdDev	0.168	0.015	0.177	0.015	0.313	0.066
Range	0.624	0.083	0.643	0.079	1.095	0.238
Minimum	1.195	0.889	1.195	0.907	1.267	0.983
Maximum	1.820	0.971	1.839	0.986	2.363	1.221

> VERTICALL

	G_B	T_B	S_B
Medn	0.945	0.935	1.091
Mean	0.944	0.934	1.096
SD	0.015	0.015	0.066
IQR	0.020	0.018	0.108
Rng	0.079	0.083	0.238
Min	0.907	0.889	0.983
Max	0.986	0.971	1.221

> VERTICAL7

	G_B	T_B	S_B
Medn	0.934	0.922	1.103
Mean	0.936	0.923	1.106
SD	0.015	0.016	0.075
IQR	0.022	0.021	0.115
Rng	0.058	0.068	0.238
Min	0.907	0.889	0.983
Max	0.964	0.957	1.221

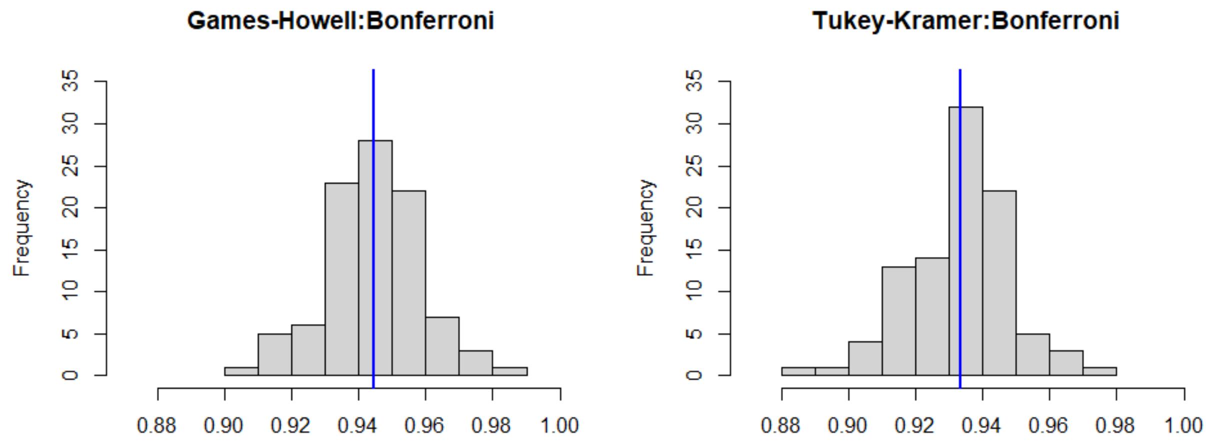
> VERTICAL8

	G_B	T_B	S_B
Medn	0.946	0.935	1.097
Mean	0.948	0.938	1.096
SD	0.012	0.011	0.063
IQR	0.016	0.013	0.093
Rng	0.048	0.048	0.191
Min	0.920	0.918	1.000
Max	0.969	0.966	1.191

> VERTICAL9

	G_B	T_B	S_B
Medn	0.947	0.939	1.084
Mean	0.950	0.940	1.085
SD	0.013	0.013	0.059
IQR	0.014	0.011	0.078
Rng	0.055	0.054	0.186
Min	0.931	0.917	0.991
Max	0.986	0.971	1.177

Figure 1.



Based on the variation in these sample size ratios across the 96 conditions from Phase 1, some percentage close to 94-95% of the Bonferroni-adjusted sample size will provide the Games-Howell MCP with sufficient power at the levels tested (93-94% for Tukey). However, some conditions required higher than 94-95% of the Bonferroni-adjusted sample size power (and some required lower). We explored these results in a second Monte Carlo experiment (Phase 2). We also extrapolated from the Phase 1 results to also test this 94-95% rule with seven, eight, and nine groups.

Phase 2 Methods and Data Sources

Phase 2 was created to verify the power rates for the MCPs when the sample sizes were calculated using the method developed in Phase 1. We ran power and sample sizes analyses using three to nine groups for power rates of .70, .80, and .90 and Cohen's *d* effect sizes from 0.2 to 0.9. Normally distributed data was generated to fit the given conditions for each simulation. That is, all data were standardized with variances of 1.0 and with all group means of 0.0 except for the one group set to the Cohen's *d* values from Phase 1. The R pwr package was used to find the Bonferroni-adjusted sample size.

All power and sample size analyses were performed using a .05 level of significance. For three to eight groups, 20,000 replications were performed for the Monte Carlo simulations, but only 10,000 replications were performed for nine groups due to the much more extreme time requirement we encountered for nine groups. We varied the Tukey and Games-Howell sample sizes as a percentage of the Bonferroni-adjusted sample size from 93% to 96%. Therefore, Phase 2 also had 96 conditions (four percentages, eight effects, three power levels).

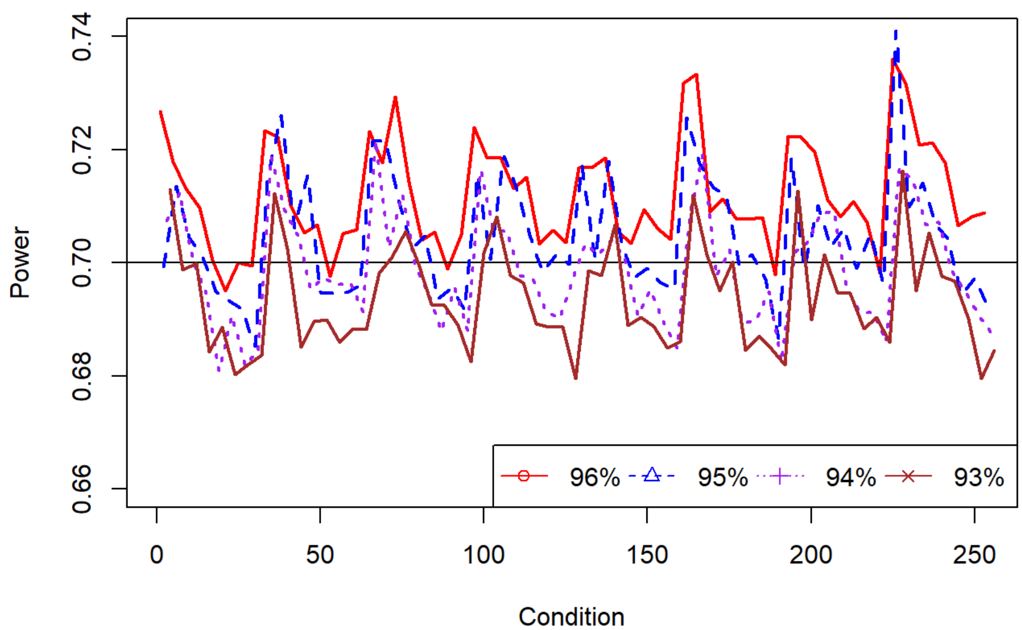
Phase 2 Results

Because many researchers use the approach of testing the homogeneity of variances assumption and then choosing Fisher *F* and Tukey if the assumption met but Welch *F* and Games-Howell of the assumption violated, we believe that choosing sample sizes based on Games-Howell is most appropriate and planning sample size based on Games-Howell makes sense. Further, Zimmerman (2004) and others have recommended performing appropriately robust tests without first testing the homogeneity assumption; in this case, Games-Howell would be the more appropriately robust choice over Tukey.

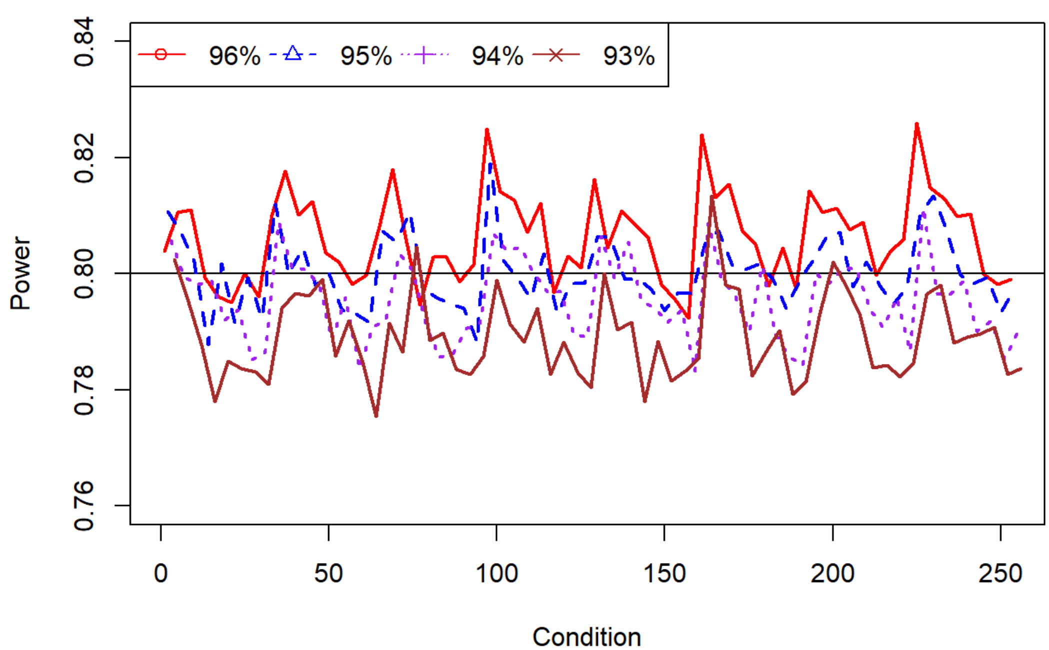
Figure 2 shows that at power of .80, using sample sizes for Games-Howell that are 96% (for Tukey at 95%) of those required for the Bonferroni-adjusted t test will generally round to the desired power. Further review suggests, however, that for Games-Howell, using 96% of the Bonferroni-adjusted sample size works well for conditions that require larger sample sizes (e.g., smaller effect sizes and more groups), 95% generally works well for medium conditions, and 94% generally works well for conditions that require smaller sample sizes to achieve power of 80% (e.g., larger effect sizes and fewer groups).

Figure 2. Power from several Games-Howell sample sizes (Conditions represent combinations of number of groups and effect sizes).

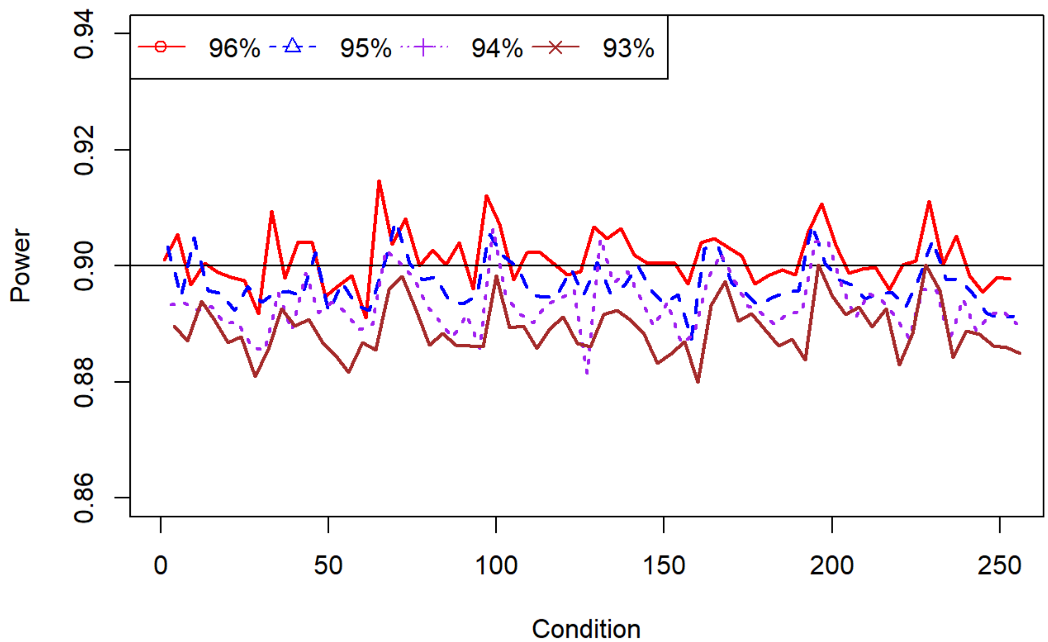
(a) Power = .70 Alpha = .05



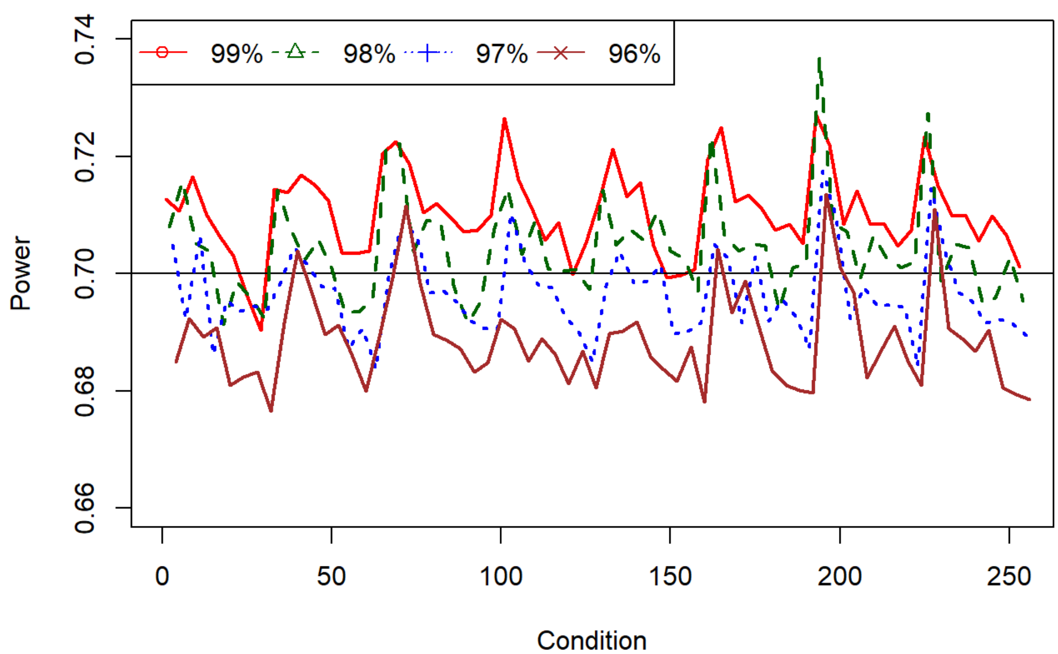
(b) Power = .80 Alpha = .05



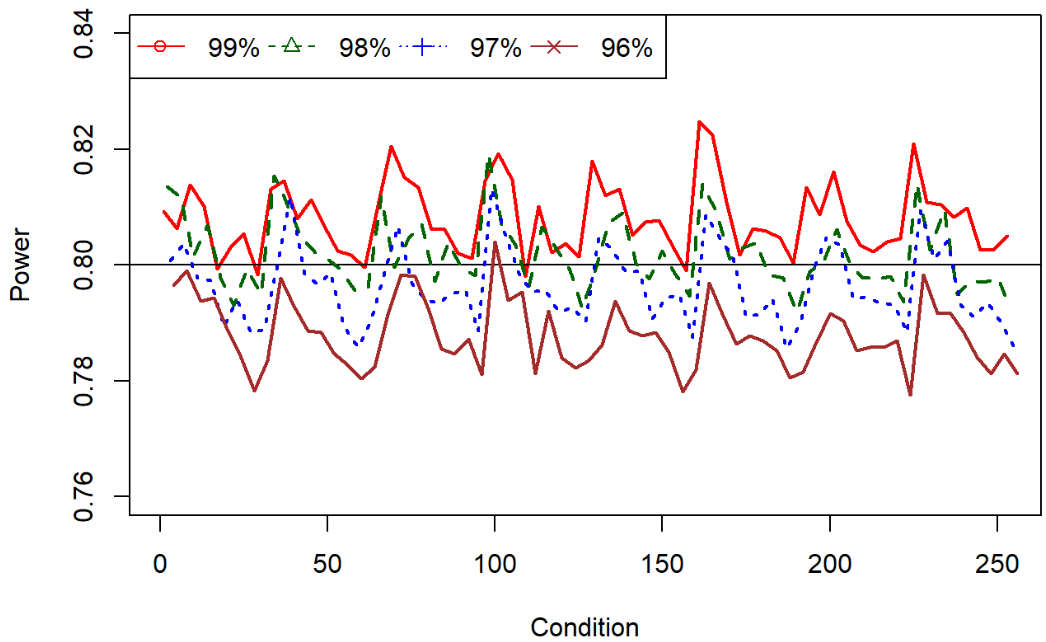
(c) Power = .90 Alpha = .05



(d) Power = .70 Alpha = .01



(e) Power = .80 Alpha = .01



(f) Power = .90 Alpha = .01

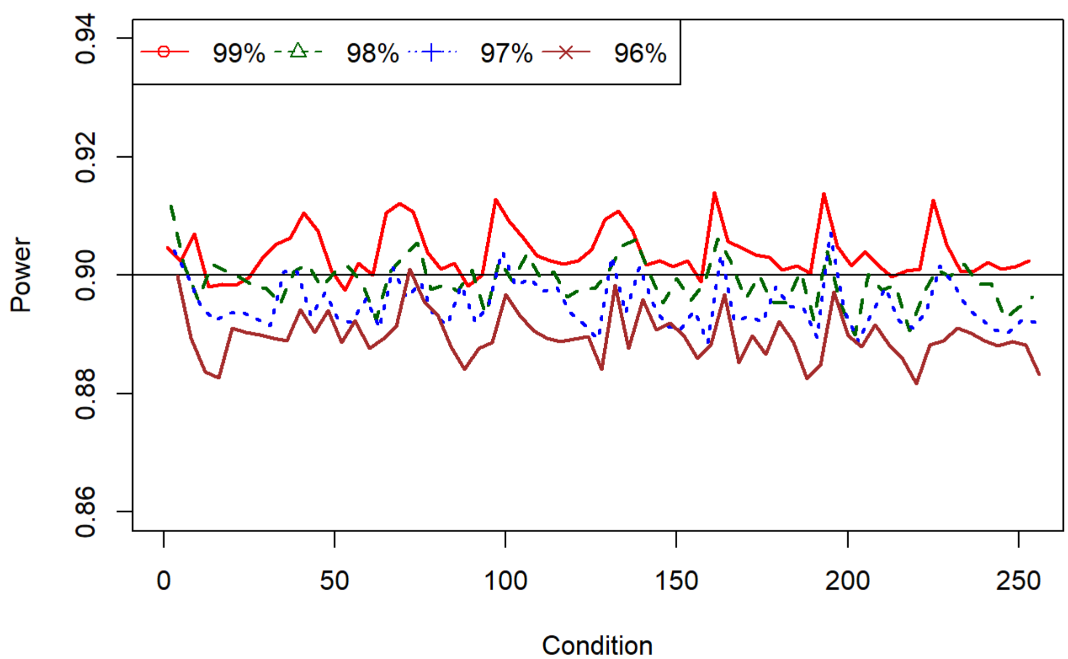
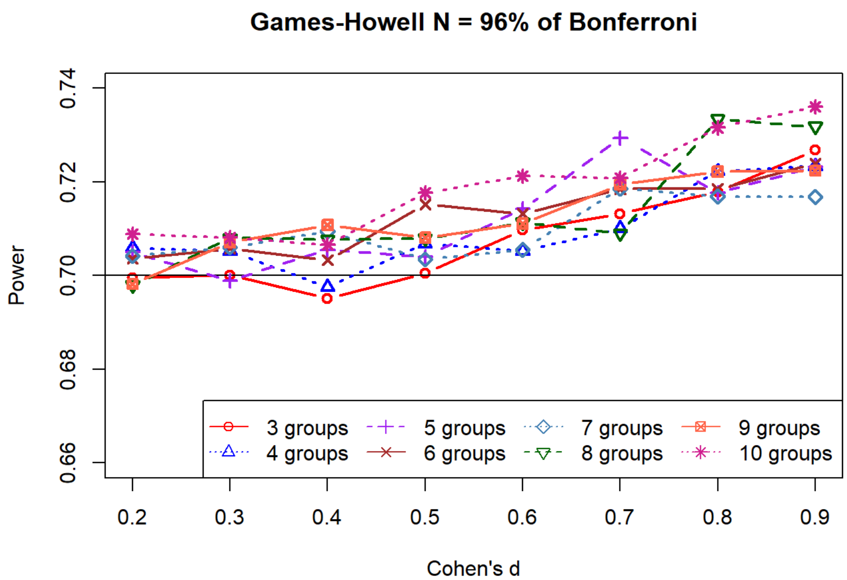
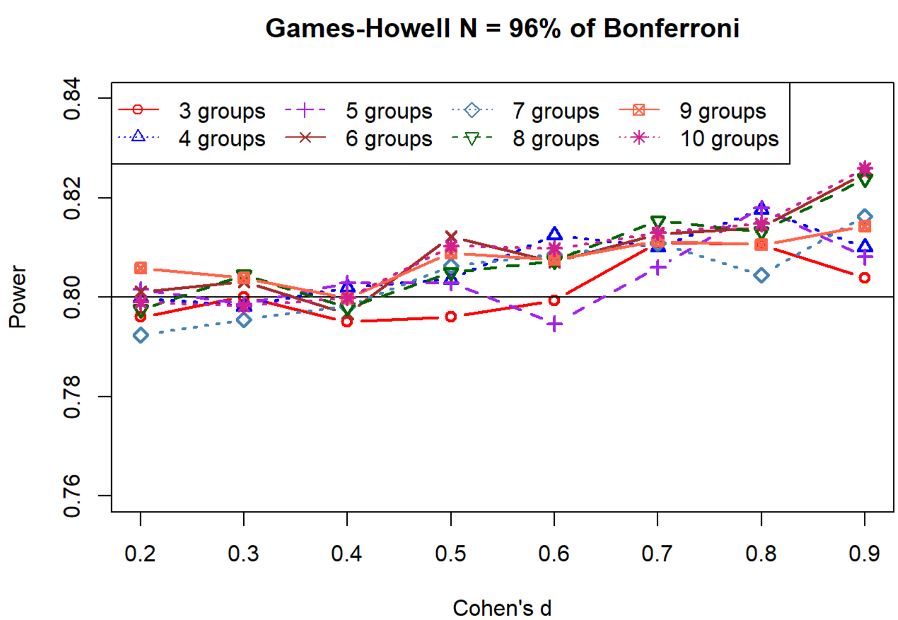


Figure 3. Percentage of Bonferroni that most consistently reaches desired power level across both groups and effect sizes

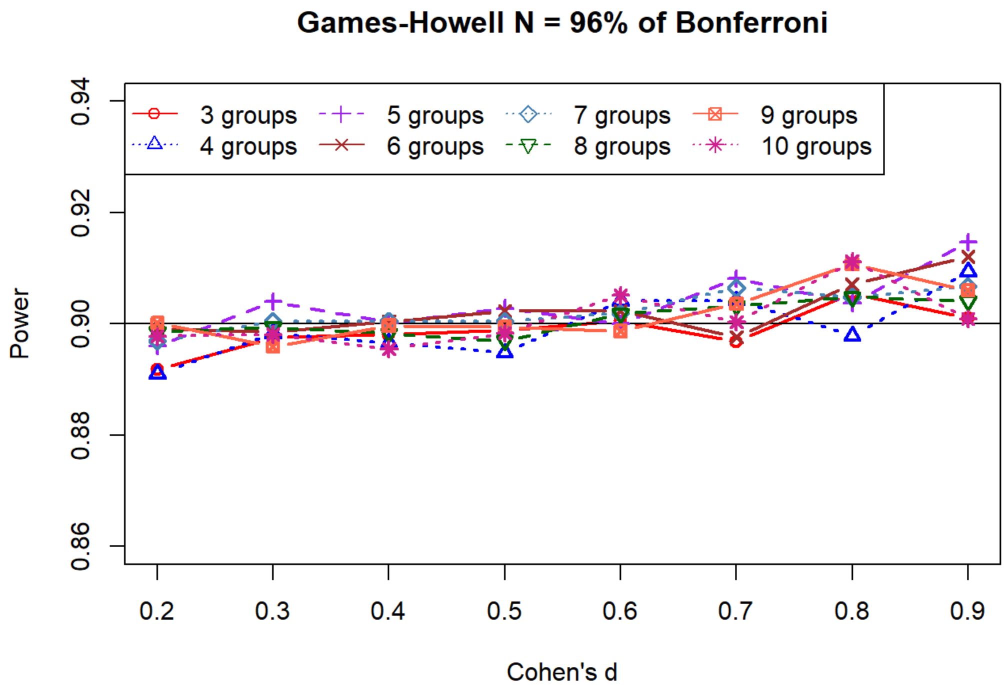
(a) Power = .70 Alpha = .05



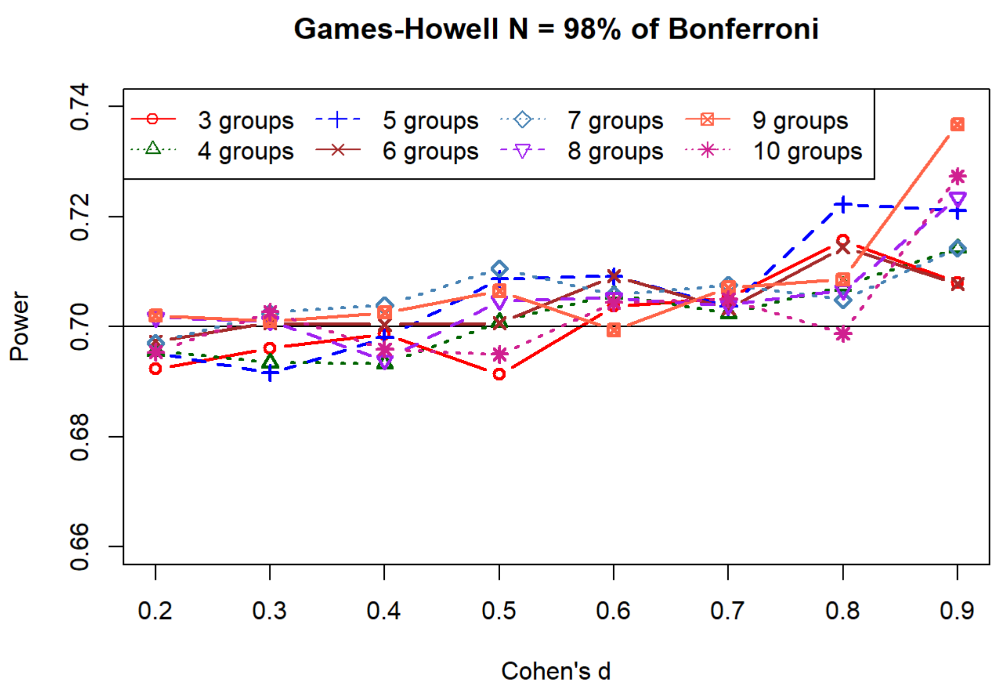
(b) Power = .80 Alpha = .05



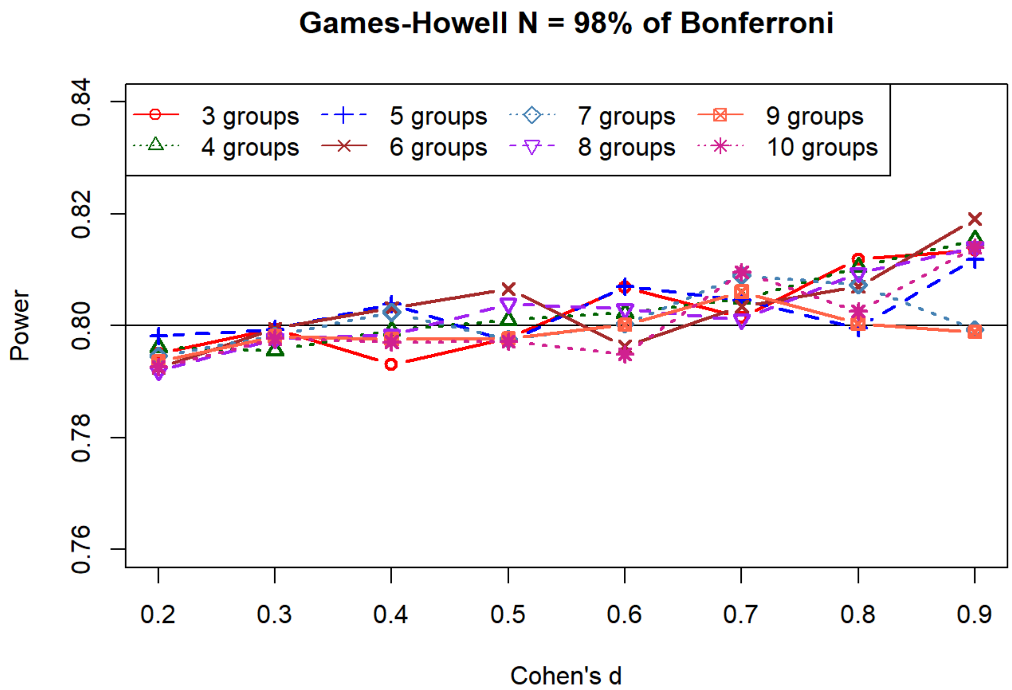
(c) Power = .90 Alpha = .05



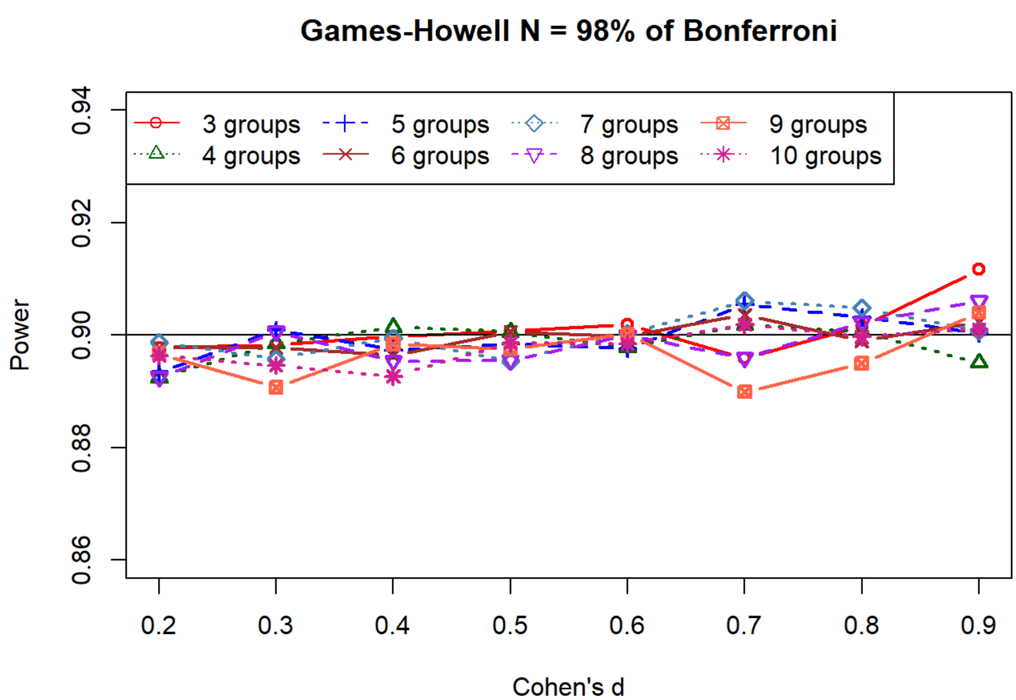
(d) Power = .70 Alpha = .01



(e) Power = .80 Alpha = .01



(f) Power = .90 Alpha = .01



Conclusions

The results provided by this study suggest that using the Bonferroni-adjusted independent t test sample sizes when determining the sample sizes for Tukey and especially Games-Howell MCPs works very well. Specifically, when specific-pair power is desired based on the comparison-of-most-interest or based on the comparison with the smallest expected effect size, then using roughly 95% of the Bonferroni-adjusted sample size generally provides the desired power level with alpha of .05. Because expected effect sizes are usually based only on a researcher's best estimate, they are not precise when determining sample sizes before collecting data. Therefore, we believe that using 95% of the Bonferroni-adjusted sample size is a reasonable recommendation for researchers who wish to determine sample sizes for Games-Howell multiple comparisons. However, this percentage can be adapted slightly for conditions that generally require smaller or larger sample sizes, by using 94% for the former and 96% for the latter. Preliminary results have shown that the same approach works for alpha of .01 but the percentage needs to be slightly higher (e.g., 96% to 98%), due to the larger sample sizes generally required for alpha of .01.

We believe the specific-pair approach to sample sizes that will be needed during the testing of multiple comparisons will be beneficial to many researchers, especially when we recognize that sample sizes required for the omnibus ANOVA do not guarantee desired power for the MCPs almost always used by researchers following a significant ANOVA. Because it will be relatively easy for researchers to obtain the Bonferroni-adjusted sample sizes using existing power programs (e.g., G*Power, SPSS, jamovi, R pwr package), we believe this approach to be a strong solution to the determining MCP sample sizes in exploratory ANOVA.

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

Brooks, G. P., An, Q., Li. Y., & Johanson, G. A. (2023). For post hoc's sake: Determining sample size for Tukey multiple comparisons in 4-Group ANOVA. *General Linear Model Journal, 47*(1), 1-14. DOI:10.31523/glmj.047001.001

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.

Hsu, J. (1996). *Multiple comparisons: Theory and Methods.* Chapman and Hall/CRC.

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