Post Hoc Power: Tables and Commentary

Russell V. Lenth

July, 2007
The University of Iowa
Department of Statistics and Actuarial Science
Technical Report No. 378

Abstract

Post hoc power is the retrospective power of an observed effect based on the sample size and parameter estimates derived from a given data set. Many scientists recommend using post hoc power as a follow-up analysis, especially if a finding is nonsignificant. This article presents tables of post hoc power for common t and F tests. These tables make it explicitly clear that for a given significance level, post hoc power depends only on the P value and the degrees of freedom. It is hoped that this article will lead to greater understanding of what post hoc power is—and is not. We also present a "grand unified formula" for post hoc power based on a reformulation of the problem, and a discussion of alternative views.

Key words: Post hoc power, Observed power, P value, Grand unified formula

1 Introduction

Power analysis has received an increasing amount of attention in the social-science literature (e.g., Cohen, 1988; Bausell and Li, 2002; Murphy and Myors, 2004). Used prospectively, it is used to determine an adequate sample size for a planned study (see, for example, Kraemer and Thiemann, 1987); for a stated effect size and significance level for a statistical test, one finds the sample size for which the power of the test will achieve a specified value.

Many studies are not planned with such a prospective power calculation, however; and there is substantial evidence (e.g., Mone et al., 1996; Maxwell, 2004) that many published studies in the social sciences are under-powered. Perhaps in response to this, some researchers (e.g., Fagley, 1985; Hallahan and Rosenthal, 1996; Onwuegbuzie and Leech, 2004) recommend that power be computed retrospectively. There are differing approaches to retrospective power, but the one of interest in this article is a power calculation based on the observed value of the effect size, as well as other auxiliary quantities such as the error standard deviation, while the significance level of the test is

held at a specified value. We will refer to such power calculations as "post hoc power" (PHP). Advocates of PHP recommend its use especially when a statistically nonsignificant result is obtained. The thinking here is that such a lack of significance could be due either to low power or to a truly small effect; if the post hoc power is found to be high, then the argument is made that the nonsignificance must then be due to a small effect size.

There is substantial literature, much of it outside of the social sciences (e.g., Goodman and Berlin, 1994; Zumbo and Hubley, 1998; Levine and Ensom, 2001; Hoenig and Heisey, 2001), that takes an opposing view to PHP practices. Lenth (2001) points out that PHP is simply a function of the P value of the test, and thus adds no new information. Yuan and Maxwell (2005) show that PHP does not necessarily provide an accurate estimate of true power. Hoenig and Heisey (2001) discuss several misconceptions connected with retrospective power. Among other things, they demonstrate that when a test is nonsignificant, then the higher the PHP, the more evidence there is *against* the null hypothesis. They also point out that, in lieu of PHP, a correct and effective way to establish that an effect is small is to use an equivalence test (Schuirmann, 1987).

In this article, we derive and present new tables that directly give exact PHP for all standard scenarios involving t tests (Section 2) and F tests (Section 3). (The PHP of certain z tests and χ^2 tests can also be obtained as limiting cases.) All that is needed to obtain PHP in these settings is the significance level, the P value of the test, and the degrees of freedom. If one desires a PHP calculation, this is obviously a convenient resource for obtaining exact power with very little effort; however, the broader goal is to demonstrate explicitly what PHP is, and what it is not. In Section 4, we present a slight reformulation of the PHP problem that leads to a "grand unified formula" for post hoc power that is universal to all tests and is a simple head calculation. The results are discussed in Section 5, along with possible alternative practices regarding retrospective power.

2 *t* tests

Table 1 may be used to obtain the post hoc power (PHP) for most common one-and two-tailed t tests, when the significance level is $\alpha = .05$. The only required information (beyond α) is the P value of the test and the degrees of freedom. Computational details are provided later in this section; for now, here is an illustration based on an example in Hallahan and Rosenthal (1996). They discuss the results of a hypothetical study where a new treatment is tested to see if it improves cognitive functioning of stroke victims. There are 20 patients in the control group and 20 in the treatment group, and the observed difference between the groups is .4 standard deviations—somewhat short of a "medium" effect on the scale proposed by Cohen (1988)—with a P value of .225 (two-sample pooled t test, two-tailed). In this case, we have $\nu = 38$ degrees of freedom. Referring to the bottom half of Table 1 (for 2-sided tests) and linearly interpolating, we obtain a post hoc power of about .234 (the exact value, using the algorithm used to produce Table 1, is .2251.) This agrees with the value of .23 reported in the article.

We briefly discuss some patterns in these tables. First, PHP is a decreasing function of

Table 1: Post hoc power of a t test when the significance level is $\alpha = .05$. It depends on the P value, the degrees of freedom ν, and whether it is one- or two-tailed. Post hoc power of a z test may be obtained using the entries for $\nu = \infty$.

		P value of test						
Alternative	ν	0.001	0.01	0.05	0.1	0.25	0.5	0.75
One-tailed	1	1.0000	1.0000	0.6767	0.3698	0.1348	0.0500	0.0105
	2	1.0000	0.9910	0.5996	0.3571	0.1434	0.0500	0.0118
	5	0.9995	0.8899	0.5365	0.3565	0.1557	0.0500	0.0112
	10	0.9860	0.8225	0.5174	0.3573	0.1607	0.0500	0.0107
	20	0.9627	0.7870	0.5084	0.3578	0.1633	0.0500	0.0105
	50	0.9420	0.7660	0.5033	0.3580	0.1649	0.0500	0.0103
	200	0.9300	0.7556	0.5008	0.3582	0.1657	0.0500	0.0102
	1000	0.9267	0.7529	0.5002	0.3582	0.1659	0.0500	0.0102
	∞	0.9258	0.7522	0.5000	0.3582	0.1659	0.0500	0.0102
Two-tailed	1	1.0000	1.0000	0.6812	0.3797	0.1506	0.0730	0.0542
	2	1.0000	0.9922	0.6147	0.3731	0.1619	0.0804	0.0562
	5	0.9996	0.8919	0.5446	0.3727	0.1864	0.0918	0.0589
	10	0.9844	0.8145	0.5210	0.3744	0.1978	0.0973	0.0602
	20	0.9553	0.7723	0.5102	0.3754	0.2038	0.1003	0.0609
	50	0.9290	0.7473	0.5040	0.3761	0.2075	0.1022	0.0614
	200	0.9137	0.7351	0.5010	0.3764	0.2094	0.1032	0.0616
	1000	0.9094	0.7318	0.5002	0.3765	0.2099	0.1035	0.0617
	∞	0.9083	0.7310	0.5000	0.3765	0.2100	0.1035	0.0617

P value, for any number of degrees of freedom and alternative. In general, except for very small degrees of freedom, the power of a marginally significant test ($P = \alpha = .05$) is around one half, with the two-tailed powers generally higher than the one-tailed results. If the test is significant, the power is higher than .5; and when the test is nonsignificant, the power is usually less than .5. Thus, it is an empty question whether the PHP is high when significance is not achieved.

2.1 Derivation of the tables

Consider the null hypothesis $H_0 : \theta = \theta_0$, where θ is some parameter and θ_0 is a specified null value (often zero). We have available an estimator $\hat{\theta}$, and the t statistic has the form

$$t = \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})} \tag{1}$$

where $se(\hat{\theta})$ is an estimate of the standard error of $\hat{\theta}$ when H_0 is true. Assume that:

1. For all θ, $\hat{\theta}$ is normally distributed with mean θ; its standard deviation will be denoted τ.

2. For all θ, $\nu \cdot se(\hat{\theta})^2/\tau^2$ has a χ^2 distribution with ν degrees of freedom. The value of ν is known.

3. $\hat{\theta}$ and $se(\hat{\theta})$ are independent.

These conditions hold for most common t-test settings, such as a one-sample test of a mean, pooled or paired comparisons of two means, and tests of regression coefficients under standard homogeneity assumptions.

Let us re-write (1) in the form

$$t = \frac{[(\hat{\theta} - \theta)/\tau] \; + \; [(\theta - \theta_0)/\tau]}{se(\hat{\theta})/\tau} = \frac{Z + \delta}{\sqrt{Q/\nu}} \tag{2}$$

where $\delta = (\theta - \theta_0)/\tau$. According to the stated assumptions, Z and Q are independent, Z is standard normal, and Q is χ^2 with ν degrees of freedom. This characterizes the noncentral t distribution with ν degrees of freedom and noncentrality parameter δ. (See, for example, Hogg et al., 2005, page 442). The power of the test is then defined as $P(t \in R_{H_1,\alpha})$, where $R_{H_1,\alpha}$ is the set of t values for which H_0 is rejected, based on the stated alternative H_1 and significance level α.

Notice that the form of $\delta = (\theta - \theta_0)/\tau$ is exactly that of the t statistic, with population values substituted in place of $\hat{\theta}$ and $se(\hat{\theta})$. In calculating PHP, we substitute the observed values of $\hat{\theta}$ and the observed error standard deviation (and thus the observed $se(\hat{\theta})$) for their population counterparts; thus, the noncentrality parameter used in PHP is $\hat{\delta} = t$, the observed t statistic itself. If one is given only the P value and the degrees of freedom, the inverse of the t distribution may be used to obtain the observed t statistic (or its absolute value, in the case of the two-tailed test), hence the noncentrality parameter $\hat{\delta}$, hence the post hoc power. Table 1 is computed using this process. Computations were performed in the R statistical package (R Development Core Team, 2006), using its built-in functions `qt` and `pt` (percentiles and cumulative probabilities of the central or noncentral t distribution).

Post hoc power of certain z tests can be obtained from the limiting case when $\nu \to \infty$. This can be verified by noting that the z statistic has the same form as (1) with $se(\hat{\theta})$ set to its known value τ. Then the denominator in (2) reduces to 1. However, keep in mind the underlying condition in our derivation that the standard error of $\hat{\theta}$ is τ regardless of the true value of θ; this condition does *not* hold in z tests involving proportions, because the standard error of a proportion depends on the value of the proportion itself.

3 *F* tests

Table 2 provides PHP values for a variety of fixed-effect F tests such as those obtained in the analysis of linear models with homogeneous-variance assumptions. Given a significance level of $\alpha = .05$ (the only case covered in the tables), the only other information needed to obtain PHP is the P value and the numerator and denominator degrees of freedom (ν_1 and ν_2 respectively). For example, suppose that we have data from an experiment where scores were measured on 40 children randomly assigned to 5

Table 2: Post hoc power of a fixed-effects F test when the significance level is $\alpha = .05$. PHP depends on the P value of the test and the degrees of freedom for the numerator (ν_1) and the denominator (ν_2). Post hoc power of a χ^2 test with ν_1 degrees of freedom may be obtained using the entries for $\nu_2 = \infty$. Post hoc power for $\nu_1 = 1$ may be obtained from the two-tailed t-test results in Table 1, with $\nu = \nu_2$.

		P value of test						
ν_1	ν_2	0.001	0.01	0.05	0.1	0.25	0.5	0.75
2	1	1.0000	1.0000	0.6827	0.3829	0.1587	0.0818	0.0593
	2	1.0000	0.9933	0.6326	0.3943	0.1823	0.0963	0.0657
	5	0.9998	0.9157	0.5951	0.4249	0.2320	0.1248	0.0774
	10	0.9899	0.8527	0.5865	0.4444	0.2615	0.1427	0.0850
	20	0.9668	0.8166	0.5842	0.4563	0.2794	0.1542	0.0899
	50	0.9436	0.7949	0.5835	0.4642	0.2913	0.1621	0.0934
	200	0.9296	0.7843	0.5834	0.4683	0.2976	0.1663	0.0952
	1000	0.9257	0.7815	0.5834	0.4694	0.2993	0.1675	0.0958
	∞	0.9247	0.7808	0.5834	0.4697	0.2997	0.1678	0.0959
3	1	1.0000	1.0000	0.6831	0.3837	0.1607	0.0846	0.0615
	2	1.0000	0.9936	0.6386	0.4015	0.1897	0.1028	0.0708
	5	0.9999	0.9266	0.6205	0.4514	0.2555	0.1431	0.0904
	10	0.9926	0.8747	0.6256	0.4864	0.3006	0.1730	0.1054
	20	0.9741	0.8454	0.6324	0.5095	0.3309	0.1944	0.1164
	50	0.9545	0.8281	0.6381	0.5253	0.3521	0.2101	0.1248
	200	0.9424	0.8198	0.6415	0.5338	0.3636	0.2189	0.1295
	1000	0.9389	0.8176	0.6425	0.5361	0.3668	0.2213	0.1309
	∞	0.9381	0.8171	0.6427	0.5367	0.3676	0.2220	0.1312
4	1	1.0000	1.0000	0.6832	0.3841	0.1615	0.0859	0.0627
	2	1.0000	0.9938	0.6416	0.4051	0.1934	0.1063	0.0738
	5	0.9999	0.9329	0.6363	0.4681	0.2705	0.1552	0.0995
	10	0.9942	0.8893	0.6528	0.5162	0.3289	0.1957	0.1218
	20	0.9792	0.8662	0.6683	0.5496	0.3709	0.2272	0.1398
	50	0.9627	0.8533	0.6804	0.5734	0.4018	0.2517	0.1543
	200	0.9525	0.8475	0.6874	0.5864	0.4190	0.2660	0.1629
	1000	0.9495	0.8460	0.6894	0.5900	0.4238	0.2700	0.1654
	∞	0.9488	0.8456	0.6899	0.5909	0.4250	0.2710	0.1661
10	1	1.0000	1.0000	0.6835	0.3847	0.1629	0.0880	0.0648
	2	1.0000	0.9940	0.6469	0.4117	0.2004	0.1130	0.0799
	5	0.9999	0.9463	0.6731	0.5079	0.3071	0.1855	0.1242
	10	0.9974	0.9266	0.7290	0.6018	0.4140	0.2679	0.1783
	20	0.9915	0.9256	0.7806	0.6807	0.5111	0.3524	0.2386
	50	0.9859	0.9310	0.8225	0.7435	0.5947	0.4336	0.3015
	200	0.9830	0.9359	0.8467	0.7796	0.6457	0.4875	0.3465
	1000	0.9822	0.9374	0.8534	0.7897	0.6603	0.5037	0.3605
	∞	0.9821	0.9378	0.8551	0.7922	0.6641	0.5080	0.3642

groups of 8 each, and the groups represent different learning conditions. We ran a one-way analysis of variance (ANOVA) to test the null hypothesis that there is no difference among the mean scores of these groups, and it was found that the P value was about .75. Since the degrees of freedom are $(\nu_1 = 4, \nu_2 = 35)$, we find in Table 2 that the PHP is somewhere between .14 and .15.

Table 2 does not cover the case where there is 1 numerator degree of freedom; this is because an F test with one numerator degree of freedom is equivalent to a two-sided t test, with $t^2 = F$. Hence, PHPs for that case can be found by referring to Table 1.

Examining the table broadly, we notice that, all other things being equal, the PHP increases with the numerator degrees of freedom. Also, as before, PHP is a decreasing function of the P value. In marginally significant cases ($P = .05$), the power is greater than .50, often by quite a bit. There are even cases with $P = .1, .25$, and even .5 where PHP exceeds .50. This is evidence of the fact that PHP is positively biased for F tests, as is shown later in this section.

There is another, quite different, situation where F tests are used to compare two independent sample variances, or to test a random effect in an ANOVA model. Table 3 provides post-hoc power values for such random-effects F tests (only a right-tailed alternative is covered). Again, the required information to use the table are the P value and the degrees of freedom. The last section of the table is for equal degrees of freedom $\nu_1 = \nu_2$, which is the case when we compare the variances of two equal-sized samples. The values in this table are quite different from those in Table 2. When $P = \alpha = .05$, the PHP is exactly .5 whenever $\nu_1 = \nu_2$, and greater or less than that when $\nu_1 > \nu_2$ or $\nu_1 < \nu_2$. We do not have the bias issue that we had for fixed effects, because the inputs to the PHP calculation are in fact two independent unbiased estimates of their respective variances.

3.1 Derivation for the fixed-effects case

Our derivation of the results needed for Table 2 uses an assumption that the F statistic is a ratio of quadratic forms, such as is the case in linear models. Let $\mathbf{y}$ be a random vector of length n having a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The F statistic has the form

$$F = \frac{\mathbf{y}'\mathbf{A}_1\mathbf{y} \;/\; \nu_1}{\mathbf{y}'\mathbf{A}_2\mathbf{y} \;/\; \nu_2} \tag{3}$$

where $\mathbf{A}_1$ and $\mathbf{A}_2$ are $n \times n$ idempotent matrices, $\nu_1 = \text{rank}(\mathbf{A}_1) = \text{tr}(\mathbf{A}_1)$, and $\nu_2 = \text{rank}(\mathbf{A}_2) = \text{tr}(\mathbf{A}_2)$. Referring to standard results in linear models (e.g., Hogg et al., 2005, Sections 9.8–9.9), we can establish that F has a noncentral F distribution provided that the following conditions hold:

1. $\mathbf{A}_1\boldsymbol{\Sigma}\mathbf{A}_2 = \mathbf{0}$ (this ensures the numerator and denominator are independent).
2. $\boldsymbol{\mu}'\mathbf{A}_2\boldsymbol{\mu} = 0$ (i.e., the noncentrality parameter of the denominator is zero).
3. $\text{tr}(\mathbf{A}_1\boldsymbol{\Sigma})/\nu_1 = \text{tr}(\mathbf{A}_2\boldsymbol{\Sigma})/\nu_2$. Since the expectation of $\mathbf{y}'\mathbf{A}_i\mathbf{y}$ is equal to $\boldsymbol{\mu}'\mathbf{A}_i\boldsymbol{\mu} + \text{tr}(\mathbf{A}_i\boldsymbol{\Sigma})$, this condition states that the expected mean squares of the numerator and denominator differ only by λ/ν_1.

Table 3: Post hoc power of a random-effects F test with a right-tailed alternative, when the significance level is $\alpha = .05$. The PHP depends on the P value of the test and the degrees of freedom for the numerator (ν_1) and the denominator (ν_2).

		P value of test						
ν_1	ν_2	0.001	0.01	0.05	0.1	0.25	0.5	0.75
1	1	0.9873	0.8746	0.5000	0.2936	0.1195	0.0500	0.0207
	2	0.9042	0.7069	0.4226	0.2785	0.1154	0.0342	0.0071
	5	0.7236	0.5518	0.3632	0.2581	0.1051	0.0166	0.0006
	10	0.6376	0.4981	0.3409	0.2471	0.0981	0.0098	0.0000
	20	0.5939	0.4720	0.3293	0.2406	0.0936	0.0065	0.0000
	50	0.5682	0.4567	0.3221	0.2364	0.0906	0.0047	0.0000
	200	0.5556	0.4492	0.3185	0.2342	0.0890	0.0039	0.0000
	1000	0.5522	0.4472	0.3176	0.2336	0.0885	0.0037	0.0000
	∞	0.5514	0.4467	0.3173	0.2334	0.0884	0.0037	0.0000
2	1	0.9996	0.9623	0.5774	0.3322	0.1358	0.0612	0.0312
	2	0.9813	0.8390	0.5000	0.3214	0.1364	0.0500	0.0172
	5	0.8597	0.6691	0.4312	0.3029	0.1318	0.0333	0.0046
	10	0.7649	0.5973	0.4019	0.2904	0.1259	0.0243	0.0013
	20	0.7083	0.5599	0.3855	0.2821	0.1213	0.0190	0.0004
	50	0.6724	0.5371	0.3751	0.2764	0.1178	0.0156	0.0001
	200	0.6542	0.5256	0.3697	0.2733	0.1159	0.0139	0.0000
	1000	0.6493	0.5225	0.3682	0.2725	0.1154	0.0134	0.0000
	∞	0.6481	0.5218	0.3679	0.2723	0.1152	0.0133	0.0000
5	1	1.0000	0.9959	0.6368	0.3608	0.1475	0.0688	0.0384
	2	0.9995	0.9389	0.5688	0.3562	0.1516	0.0620	0.0274
	5	0.9630	0.7926	0.5000	0.3433	0.1529	0.0500	0.0135
	10	0.8915	0.7085	0.4651	0.3309	0.1492	0.0412	0.0069
	20	0.8295	0.6572	0.4430	0.3210	0.1450	0.0347	0.0036
	50	0.7823	0.6228	0.4275	0.3132	0.1411	0.0299	0.0018
	200	0.7557	0.6043	0.4189	0.3086	0.1387	0.0271	0.0011
	1000	0.7483	0.5993	0.4165	0.3072	0.1379	0.0263	0.0010
	∞	0.7464	0.5980	0.4159	0.3069	0.1378	0.0261	0.0009
$\nu_1 = \nu_2$	1	0.9873	0.8746	0.5000	0.2936	0.1195	0.0500	0.0208
	2	0.9813	0.8390	0.5000	0.3214	0.1364	0.0500	0.0172
	5	0.9630	0.7926	0.5000	0.3433	0.1529	0.0500	0.0135
	10	0.9480	0.7728	0.5000	0.3509	0.1593	0.0500	0.0119
	20	0.9378	0.7626	0.5000	0.3546	0.1626	0.0500	0.0111
	50	0.9308	0.7564	0.5000	0.3568	0.1646	0.0500	0.0105
	200	0.9271	0.7533	0.5000	0.3578	0.1656	0.0500	0.0103
	1000	0.9261	0.7524	0.5000	0.3581	0.1659	0.0500	0.0102

While the elements of $\mathbf{\Sigma}$ are assumed unknown, we assume that enough is known about its structure (e.g., diagonal or compound-symmetric) that these conditions can be verified.

The distribution of F has degrees of freedom (ν_1, ν_2) and noncentrality parameter $\lambda = \boldsymbol{\mu}'\mathbf{A}_1\boldsymbol{\mu}/\tau^2$, where $\tau^2 = \text{tr}(\mathbf{A}_2\mathbf{\Sigma})/\nu_2$, the expected value of the denominator of F. The hypotheses under test are $H_0 : \lambda = 0$ versus $H_1 : \lambda > 0$. The power of the test is the probability that this noncentral F random variable exceeds the $(1-\alpha)$th quantile of the central F distribution with (ν_1, ν_2) d.f. For post hoc power, we would use the observed value of the denominator as an estimate of τ^2; and estimate $\boldsymbol{\mu}'\mathbf{A}_1\boldsymbol{\mu}$ by $\mathbf{y}'\mathbf{A}_1\mathbf{y}$. Thus, the estimated noncentrality parameter for PHP is

$$\hat{\lambda} \;=\; \frac{\mathbf{y}'\mathbf{A}_1\mathbf{y}}{\mathbf{y}'\mathbf{A}_2\mathbf{y} \;/\; \nu_2} \;=\; \nu_1 \cdot F \tag{4}$$

Given ν_1, ν_2, and the P value, we can work backwards to find the value of F, then obtain $\hat{\lambda}$ and the post hoc power. Table 2 is computed using this process, using the R functions `qf` and `pf` (R Development Core Team, 2006).

Note that we can use the mean of the noncentral F distribution to show that the expectation of $\hat{\lambda}$ is $\frac{\nu_2}{\nu_2-2}(\lambda + \nu_1)$ when $\nu_2 > 2$. This shows why the PHPs in Table 2 can be so exaggerated, especially when ν_1 is large or P is large (suggesting λ is small). It also disproves a statement made in Onwuegbuzie and Leech (2004) that "observed effect size ... [is] a positively biased but consistent estimate of the effect"; it is not consistent. One may make the simple adjustment $\tilde{\lambda} = (\nu_1 - 2)\hat{\lambda}/\nu_2 - \nu_1$ to obtain an unbiased estimate of λ, and using this (when it is nonnegative) in place of $\hat{\lambda}$ substantially reduces the PHP; for example, the "bias-corrected" PHP for $\nu_1 = 10$, $\nu_2 = 50$, and $P = .25$ is .1290, compared with the value of .5947 in Table 2. Taylor and Muller (1996) provides more detailed and sophisticated approaches to dealing with bias in estimating noncentrality and power.

3.2 Derivation for the random-effects case

Derivation of results for the random-effects case is relatively simple. The F statistic has the form $F = s_1^2/s_2^2$, where s_1 and s_2 are independent random variables such that for $i = 1, 2$, $\nu_i s_i^2/\sigma_i^2$ has a χ^2 distribution with ν_i d.f. We test $H_0 : \sigma_1^2 = \sigma_2^2$ against some alternative; Table 3 only considers the right-tailed alternative $H_1 : \sigma_1^2 > \sigma_2^2$. It is clear that in the general case, $(\sigma_1^2/\sigma_2^2)F$ has a central F distribution with (ν_1, ν_2) d.f. The power of the test is the probability that this multiple of an F random variable exceeds the $(1-\alpha)$ quantile of the F distribution. To compute power retrospectively, we simply use the observed ratio $s_1^2/s_2^2 = F$ as an estimate of the ratio σ_1^2/σ_2^2. As in the fixed-effects case, we can work backwards from the P value to find the observed F value. Again, we used the R functions `qf` and `pf` to compute Table 3.

4 A grand unified formula for post hoc power

Recall that PHP is based on pre-specified hypotheses and α, but that all other parameters are estimated from the data. Going back to basics, the power of a test is the probability of rejecting the null hypothesis in favor of the alternative, computed at a specified effect size. Given the same information used in PHP calculations, we can write

$$PHP \;=\; P(\text{Reject } H_0 \mid \text{available data}) \;=\; \begin{cases} 1 & \text{if } H_0 \text{ was rejected} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

That is, when we compute post hoc power, it makes sense to use all the information available. The PHP computations described in the literature and earlier in this article all ignore an essential known fact—the outcome of the test. Certainly, (5) is easy to remember and can be applied universally to all post-hoc-power problems: a grand unified formula (GUF) of astonishing simplicity!

5 Discussion

The tables in this article demonstrate clearly that PHP is just a re-expression of the *P* value; and in fact, once one gets past 20 degrees of freedom or so (for the denominator), PHP does not even depend much on sample size for a given type of test. Thus, as a retrospective measure of the results of the current study, PHP is just elaboration, not new information.

It is of course possible to also consider the meaning of PHP as a prospective measure. That is, contemplate a future study exactly like the one we just did, with the same sample size; what is the probability of achieving statistical significance if the same effect size is observed? (This is the only situation I can think of where PHP would make sense and the GUF would not.)

If the current study resulted in statistical significance, then the results in this article show that the PHP is fairly high, indicating a good chance of achieving significance if the study is repeated; such might add credibility to our current results—but only if we actually conduct that new study.

It is in the case of a nonsignificant result that many authors recommend that PHP be calculated. We have shown that PHP tends to be low in this situation (at least, after correcting for bias in estimating the noncentrality parameter); and, viewed prospectively, that suggests that a future study of identical design and sample size is also likely to again result in nonsignificance. Moreover, don't forget that, especially in the fixed-effects *F* case, PHP over-estimates the power of that future study.

The above two paragraphs can be summarized as follows: If you were to repeat the same study, you'd probably get the same result you got this time—with some variation, of course. I think we already knew that; PHP doesn't help us understand that point any more clearly than we already did. Thus, it does not make sense prospectively, either. The bottom line is that PHP does not tell us anything we don't already know.

Researchers who advocate retrospective power calculations are motivated by the finest of objectives. Onwuegbuzie and Leech (2004), for example, make a number of

points about considering the practical importance of research results and not just whether they are statistically significant, and reinforce the recommendations in the APA publication manual (American Psychological Association, 2001, page 25) that an index of effect size should accompany the results of tests. While these are valid and important points, post hoc power does not address them in a meaningful way because, as we have demonstrated, the P value, observed effect size, and PHP are all confounded.

5.1 Power is inherently prospective

Neither PHP nor my GUF alternative add information to an analysis. Given that we have all of the information to be had from the current study, it only makes sense to focus attention on what to do in future work.

I believe that the debate over post hoc power has its roots in confusion over the foundations of hypothesis testing. Underlying that is the word "hypothesis." A hypothesis is, well, hypothetical; and it makes sense for things to be hypothetical only when you have yet to collect data.

When we learn hypothesis-testing methods, we are usually taught that the hypotheses should be formulated before collecting the data (or at least before looking at them), and also that the significance level α should be specified in advance. If we make any of those things up as we go along, then the analysis is only exploratory and we cannot use it to establish definitive scientific findings; that would have to be done in a future study. I believe that these points are well understood by most social scientists.

The ideas that hypotheses are hypothetical, and that they, and the value of α, should be formulated independently of the data, are thoroughly prospective in nature. When power analysis is done, it makes sense that effect size should also be defined prospectively. That indeed is the view taken even by some researchers who advocate retrospective power; for example, Fagley (1985) emphasizes basing it on an effect of meaningful size, rather than the observed effect size. However, even if retrospective power is computed using a meaningful effect size, it is still based on the sample size of the current study, rather than on focusing prospectively on what could be accomplished with a follow-up study of a possibly different sample size. Again, as explained by Hoenig and Heisey (2001), the appropriate way to judge the observed effect size relative to a meaningful one is to perform an equivalence test.

5.2 Effect-size considerations

Another brand of retrospective power is obtained by using the observed effect size, but considering the power of a future study with possibly a different sample size. This is a prospective view, but it falls in the category of "asterisk hunting" (Lenth, 2001); that is, we are setting a goal of collecting enough data to achieve statistical significance, without first making a judgment as to whether or not the observed effect size is of practical importance.

This is also the implicit goal of Onwuegbuzie and Leech (2004) and many other authors who emphasize identifying the *expected* effect size for use in power calculations.

Onwuegbuzie and Leech state (page 209, and citing several other references) that "a statistically nonsignificant finding can be assessed more appropriately by using the observed (true) effect" in a power calculation. I believe this to be highly misleading, as well as completely contrary to other views expressed in the same reference. First of all, the observed effect is not the true effect; it is only an estimate thereof. Second, if it is statistically nonsignificant, then the true effect is plausibly equal to zero; so either the calculated power is irrelevant, or we are so unsure of the true effect size that we can hardly rely on its value.

5.3 Conclusions

Researchers owe it to themselves to take a thoroughly prospective view of any power calculation. That involves establishing a meaningful effect size based not on anticipated results, but on scientific goals—a target effect size that is likely to be detected if it exists, and not so likely to be detected if it doesn't. Then power the study accordingly. That is the scientific way of connecting statistical significance with practical significance. Once the study is completed, power calculations do not inform us in any way as to the conclusions of the present study.

References

American Psychological Association (2001). *Publication Manual of the American Psychological Association*. American Psychological Association, Washington, DC, 5th edition.

Bausell, R. B. and Li, Y.-F. (2002). *Power Analysis for Experimental Research: A Practical Guide for the Biological, Medical and Social Sciences*. Cambridge University Press, Cambridge, UK.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Academic Press, New York, 2nd edition.

Fagley, N. S. (1985). Applied statistical power analysis and the interpretation of nonsignificant results. *Journal of Counseling Psychology*, 32:391–396.

Goodman, S. N. and Berlin, J. A. (1994). The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine*, 121(3):200–206.

Hallahan, M. and Rosenthal, R. (1996). Statistical power: Concepts, procedures, and applications. *Behavioral Research Therapy*, 34:489–499.

Hoenig, J. M. and Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations in data analysis. *The American Statistician*, 55:19–24.

Hogg, R. V., McKean, J. W., and Craig, A. T. (2005). *Introduction to Mathematical Statistics*. Pearson Prentice Hall, Upper Saddle River, NJ, 6th edition.

Kraemer, H. C. and Thiemann, S. (1987). *How Many Subjects? Statistical Power Analysis in Research*. Sage Publications, Newbury Park, CA.

Lenth, R. V. (2001). Some practical guidelines for effective sample-size determination. *The American Statistician*, 55:187–193.

Levine, M. and Ensom, M. H. (2001). Post hoc power analysis: an idea whose time has passed? *Pharmacotherapy*, 21:405–409.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, 9(2):147–163.

Mone, M. A., Mueller, G. C., and Mauland, W. (1996). The perceptions and usage of statistical power in applied psychology and management research. *Personnel Psychology*, 49(1):103–120.

Murphy, K. R. and Myors, B. (2004). *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests*. Lawrence Erlbaum Associates, Mahwah, NJ, 2nd edition.

Onwuegbuzie, A. J. and Leech, N. L. (2004). Post hoc power: A concept whose time has come. *Understanding Statistics*, 3(4):201–230.

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Schuirmann, D. (1987). A comparison of the two one-sided test procedure and the power approach for assessing the equivalence of bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15:657–680.

Taylor, D. J. and Muller, K. E. (1996). Bias in linear model power and sample size calculation due to estimating noncentrality. *Communications in Statistics: Theory and Methods*, 25:1595–1610.

Yuan, K. and Maxwell, S. E. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics*, 30(2):141–167.

Zumbo, B. D. and Hubley, A. M. (1998). A note on misconceptions concerning prospective and retrospective power. *The Statistician*, 47(2):385–388.

A R code used in calculating the tables

This appendix presents the R functions that were used to calculate the tables. Note that these functions can gracefully handle infinite degrees of freedom (R value of `Inf`).

A.1 Post hoc power for *t* tests

Arguments are the observed *P* value, degrees of freedom, a boolean flag for whether it is two-tailed (true) or one-tailed (false), and the significance level of the test.

```
retro.t = function(P, df=50, two.tailed=TRUE, alpha=.05) {
  if (two.tailed) {
    delta = qt(1 - P/2, df)
    cv = qt(1 - alpha/2, df)
    power = 1 - pt(cv, df, delta) + pt(-cv, df, delta)
  }
  else {
    delta = qt(1 - P, df)
    cv = qt(1 - alpha, df)
    power = 1 - pt(cv, df, delta)
  }
  power
}
```

A.2 Post hoc power for fixed-effects *F* tests

Arguments are the observed *P* value, numerator and denominator degrees of freedom, and the significance level of the test.

```
retro.F = function(P, numdf=1, dendf=50, alpha=.05) {
  lambda = numdf * qf(1 - P, numdf, dendf)
  cv = qf(1 - alpha, numdf, dendf)
  1 - pf(cv, numdf, dendf, lambda)
}
```

A.3 Post hoc power for random-effects *F* tests

Arguments are the same as for the fixed-effects case.

```
retro.rF = function(P, numdf=1, dendf=50, alpha=.05) {
  ratio = qf(1 - P, numdf, dendf)
  cv = qf(1 - alpha, numdf, dendf)
  1 - pf(cv / ratio, numdf, dendf)
}
```