A Power Primer

Jacob Cohen
New York University

One possible reason for the continued neglect of statistical power analysis in research in the behavioral sciences is the inaccessibility of or difficulty with the standard material. A convenient, although not comprehensive, presentation of required sample sizes is provided here. Effect-size indexes and conventional values for these are given for operationally defined small, medium, and large effects. The sample sizes necessary for .80 power to detect effects at these levels are tabled for eight standard statistical tests: (a) the difference between independent means, (b) the significance of a product–moment correlation, (c) the difference between independent *r*s, (d) the sign test, (e) the difference between independent proportions, (f) chi-square tests for goodness of fit and contingency tables, (g) one-way analysis of variance, and (h) the significance of a multiple or multiple partial correlation.

The preface to the first edition of my power handbook (Cohen, 1969) begins:

> During my first dozen years of teaching and consulting on applied statistics with behavioral scientists, I became increasingly impressed with the importance of statistical power analysis, an importance which was increased an order of magnitude by its neglect in our textbooks and curricula. The case for its importance is easily made: What behavioral scientist would view with equanimity the question of the probability that his investigation would lead to statistically significant results, i.e., its power? (p. vii)

This neglect was obvious through casual observation and had been confirmed by a power review of the 1960 volume of the *Journal of Abnormal and Social Psychology,* which found the mean power to detect medium effect sizes to be .48 (Cohen, 1962). Thus, the chance of obtaining a significant result was about that of tossing a head with a fair coin. I attributed this disregard of power to the inaccessibility of a meager and mathematically difficult literature, beginning with its origin in the work of Neyman and Pearson (1928, 1933).

The power handbook was supposed to solve the problem. It required no more background than an introductory psychological statistics course that included significance testing. The exposition was verbal–intuitive and carried largely by many worked examples drawn from across the spectrum of behavioral science.

In the ensuing two decades, the book has been through revised (1977) and second (1988) editions and has inspired dozens of power and effect-size surveys in many areas of the social and life sciences (Cohen, 1988, pp. xi–xii). During this period, there has been a spate of articles on power analysis in the social science literature, a baker's dozen of computer programs (reviewed in Goldstein, 1989), and a breakthrough into popular statistics textbooks (Cohen, 1988, pp. xii–xiii).

Sedlmeier and Gigerenzer (1989) reported a power review of the 1984 volume of the *Journal of Abnormal Psychology* (some 24 years after mine) under the title, "Do Studies of Statistical Power Have an Effect on the Power of Studies?" The answer was no. Neither their study nor the dozen other power reviews they cite (excepting those fields in which large sample sizes are used, e.g., sociology, market research) showed any material improvement in power. Thus, a quarter century has brought no increase in the probability of obtaining a significant result.

Why is this? There is no controversy among methodologists about the importance of power analysis, and there are ample accessible resources for estimating sample sizes in research planning using power analysis. My 2-decades-long expectation that methods sections in research articles in psychological journals would invariably include power analyses has not been realized. Indeed, they almost invariably do not. Of the 54 articles Sedlmeier and Gigerenzer (1989) reviewed, only 2 mentioned power, and none estimated power or necessary sample size or the population effect size they posited. In 7 of the studies, null hypotheses served as research hypotheses that were confirmed when the results were nonsignificant. Assuming a medium effect size, the median power for these tests was .25! Thus, these authors concluded that their research hypotheses of no effect were supported when they had only a .25 chance of rejecting these null hypotheses in the presence of substantial population effects.

It is not at all clear why researchers continue to ignore power analysis. The passive acceptance of this state of affairs by editors and reviewers is even more of a mystery. At least part of the reason may be the low level of consciousness about effect size: It is as if the only concern about magnitude in much psychological research is with regard to the statistical test result and its accompanying *p* value, not with regard to the psychological phenomenon under study. Sedlmeier and Gigerenzer (1989) attribute this to the accident of the historical precedence of Fi-

I am grateful to Patricia Cohen for her useful comments.

Correspondence concerning this article should be addressed to Jacob Cohen, Department of Psychology, New York University, 6 Washington Place, 5th Floor, New York, New York 10003.

Psychological Bulletin, 1992, Vol. 112, No. 1, 155–159
 0033-2909/92/$3.00

sherian theory, its hybridization with the contradictory Neyman-Pearson theory, and the apparent completeness of Fisherian null hypothesis testing: objective, mechanical, and a clear-cut go–no-go decision straddled over $p = .05$. I have suggested that the neglect of power analysis simply exemplifies the slow movement of methodological advance (Cohen, 1988, p. xiv), noting that it took some 40 years from Student's publication of the t test to its inclusion in psychological statistics textbooks (Cohen, 1990, p. 1311).

An associate editor of this journal suggests another reason: Researchers find too complicated, or do not have at hand, either my book or other reference material for power analysis. He suggests that a short rule-of-thumb treatment of necessary sample size might make a difference. Hence this article.

In this bare bones treatment, I cover only the simplest cases, the most common designs and tests, and only three levels of effect size. For readers who find this inadequate, I unhesitatingly recommend *Statistic Power Analysis for the Behavioral Sciences* (Cohen, 1988; hereafter SPABS). It covers special cases, one-sided tests, unequal sample sizes, other null hypotheses, set correlation and multivariate methods and gives substantive examples of small, medium, and large effect sizes for the various tests. It offers well over 100 worked illustrative examples and is as user friendly as I know how to make it, the technical material being relegated to an appendix.

Method

Statistical power analysis exploits the relationships among the four variables involved in statistical inference: sample size (N), significance criterion (α), population effect size (ES), and statistical power. For any statistical model, these relationships are such that each is a function of the other three. For example, in power reviews, for any given statistical test, we can determine power for given α, N, and ES. For research planning, however, it is most useful to determine the N necessary to have a specified power for given α and ES; this article addresses this use.

The Significance Criterion, α

The risk of mistakenly rejecting the null hypothesis (H_0) and thus of committing a Type I error, α, represents a policy: the maximum risk attending such a rejection. Unless otherwise stated (and it rarely is), it is taken to equal .05 (part of the Fisherian legacy; Cohen, 1990). Other values may of course be selected. For example, in studies testing several H_0s, it is recommended that $\alpha = .01$ per hypothesis in order that the experimentwise risk (i.e., the risk of any false rejections) not become too large. Also, for tests whose parameters may be either positive or negative, the α risk may be defined as two sided or one sided. The many tables in SPABS provide for both kinds, but the sample sizes provided in this note are all for two-sided tests at $\alpha = .01, .05$, and .10, the last for circumstances in which a less rigorous standard for rejection is desired, as, for example, in exploratory studies. For unreconstructed one tailers (see Cohen, 1965), the tabled sample sizes provide close approximations for one-sided tests at $\frac{1}{2}\alpha$ (e.g., the sample sizes tabled under $\alpha = .10$ may be used for one-sided tests at $\alpha = .05$).

Power

The statistical power of a significance test is the long-term probability, given the population ES, α, and N of rejecting H_0. When the ES is not equal to zero, H_0 is false, so failure to reject it also incurs an error. This is a Type II error, and for any given ES, α, and N, its probability of occurring is β. Power is thus $1 - \beta$, the probability of rejecting a false H_0.

In this treatment, the only specification for power is .80 (so $\beta = .20$), a convention proposed for general use. (SPABS provides for 11 levels of power in most of its N tables.) A materially smaller value than .80 would incur too great a risk of a Type II error. A materially larger value would result in a demand for N that is likely to exceed the investigator's resources. Taken with the conventional $\alpha = .05$, power of .80 results in a $\beta{:}\alpha$ ratio of 4:1 (.20 to .05) of the two kinds of risks. (See SPABS, pp. 53–56.)

Sample Size

In research planning, the investigator needs to know the N necessary to attain the desired power for the specified α and hypothesized ES. N increases with an increase in the power desired, a decrease in the ES, and a decrease in α. For statistical tests involving two or more groups, N as here defined is the necessary sample size for *each* group.

Effect Size

Researchers find specifying the ES the most difficult part of power analysis. As suggested above, the difficulty is at least partly due to the generally low level of consciousness of the magnitude of phenomena that characterizes much of psychology. This in turn may help explain why, despite the stricture of methodologists, significance testing is so heavily preferred to confidence interval estimation, although the wide intervals that usually result may also play a role (Cohen, 1990). However, neither the determination of power or necessary sample size can proceed without the investigator having some idea about the degree to which the H_0 is believed to be false (i.e., the ES).

In the Neyman-Pearson method of statistical inference, in addition to the specification of H_0, an alternate hypothesis (H_1) is counterpoised against H_0. The degree to which H_0 is false is indexed by the discrepancy between H_0 and H_1 and is called the ES. Each statistical test has its own ES index. All the indexes are scale free and continuous, ranging upward from zero, and for all, the H_0 is that ES = 0. For example, for testing the product–moment correlation of a sample for significance, the ES is simply the population r, so H_0 posits that $r = 0$. As another example, for testing the significance of the departure of a population proportion (P) from .50, the ES index is $g = P - .50$, so the H_0 is that $g = 0$. For the tests of the significance of the difference between independent means, correlation coefficients, and proportions, the H_0 is that the difference equals zero. Table 1 gives for each of the tests the definition of its ES index.

To convey the meaning of any given ES index, it is necessary to have some idea of its scale. To this end, I have proposed as conventions or operational definitions small, medium, and large values for each that are at least approximately consistent across the different ES indexes. My intent was that medium ES represent an effect likely to be visible to the naked eye of a careful observer. (It has since been noted in effect-size surveys that it approximates the average size of observed effects in various fields.) I set small ES to be noticeably smaller than medium but not so small as to be trivial, and I set large ES to be the same distance above medium as small was below it. Although the definitions were made subjectively, with some early minor adjustments, these conventions have been fixed since the 1977 edition of SPABS and have come into general use. Table 1 contains these values for the tests considered here.

In the present treatment, the H_1s are the ESs that operationally define small, medium, and large effects as given in Table 1. For the test of the significance of a sample r, for example, because the ES for this test is simply the alternate-hypothetical population r, small, medium, and large ESs are respectively .10, .30, and .50. The ES index for the t test of the difference between independent means is d, the difference

Table 1
ES Indexes and Their Values for Small, Medium, and Large Effects

Test	ES index	Effect size		
		Small	Medium	Large
1. m_A vs. m_B for independent means	$d = \frac{m_A - m_B}{\sigma}$	.20	.50	.80
2. Significance of product–moment r	r	.10	.30	.50
3. r_A vs. r_B for independent rs	$q = z_A - z_B$ where z = Fisher's z	.10	.30	.50
4. $P = .5$ and the sign test	$g = P - .50$	.05	.15	.25
5. P_A vs. P_B for independent proportions	$h = \phi_A - \phi_B$ where ϕ = arcsine transformation	.20	.50	.80
6. Chi-square for goodness of fit and contingency	$w = \sqrt{\sum_{i=1}^{k} \frac{(P_{1i} - P_{0i})^2}{P_{0i}}}$	.10	.30	.50
7. One-way analysis of variance	$f = \frac{\sigma_m}{\sigma}$	.10	.25	.40
8. Multiple and multiple partial correlation	$f^2 = \frac{R^2}{1 - R^2}$	.02	.15	.35

Note. ES = population effect size.

expressed in units of (i.e., divided by) the within-population standard deviation. For this test, the H_0 is that $d = 0$ and the small, medium, and large ESs (or H_1s) are d = .20, .50, and .80. Thus, an operationally defined medium difference between means is half a standard deviation; concretely, for IQ scores in which the population standard deviation is 15, a medium difference between means is 7.5 IQ points.

Statistical Tests

The tests covered here are the most common tests used in psychological research:

1. The t test for the difference between two independent means, with $df = 2(N - 1)$.

2. The t test for the significance of a product–moment correlation coefficient r, with $df = N - 2$.

3. The test for the difference between two independent rs, accomplished as a normal curve test through the Fisher z transformation of r (tabled in many statistical texts).

4. The binomial distribution or, for large samples, the normal curve (or equivalent chi-square, 1 df) test that a population proportion (P) = .50. This test is also used in the nonparametric sign test for differences between paired observations.

5. The normal curve test for the difference between two independent proportions, accomplished through the arcsine transformation ϕ (tabled in many statistical texts). The results are effectively the same when the test is made using the chi-square test with 1 degree of freedom.

6. The chi-square test for goodness of fit (one way) or association in two-way contingency tables. In Table 1, k is the number of cells and P_{0i} and P_{1i} are the null hypothetical and alternate hypothetical population proportions in cell i. (Note that w's structure is the same as chi-square's for cell sample frequencies.) For goodness-of-fit tests, the $df = k - 1$, and for contingency tables, $df = (a - 1)(b - 1)$, where a and b are the number of levels in the two variables. Table 2 provides (total) sample sizes for 1 through 6 degrees of freedom.

7. One-way analysis of variance. Assuming equal sample sizes (as we do throughout), for g groups, the F test has $df = g - 1, g(N - 1)$. The ES index is the standard deviation of the g population means divided by the common within-population standard deviation. Provision is made in Table 2 for 2 through 7 groups.

8. Multiple and multiple partial correlation. For k independent variables, the significance test is the standard F test for $df = k, N - k - 1$. The ES index, f^2, is defined for either squared multiple or squared multiple partial correlations (R^2). Table 2 provides for 2 through 8 independent variables.

Note that because all tests of population parameters that can be either positive or negative (Tests 1–5) are two-sided, their ES indexes here are absolute values.

In using the material that follows, keep in mind that the ES posited by the investigator is what he or she believes holds for the population and that the sample size that is found is conditional on the ES. Thus, if a study is planned in which the investigator believes that a population r is of medium size (ES = r = .30 from Table 1) and the t test is to be performed with two-sided $\alpha = .05$, then the power of this test is .80 if the sample size is 85 (from Table 2). If, using 85 cases, t is not significant, then

Table 2
N for Small, Medium, and Large ES at Power = .80 for α = .01, .05, and .10

	α								
	.01			.05			.10		
Test	Sm	Med	Lg	Sm	Med	Lg	Sm	Med	Lg
1. Mean dif	586	95	38	393	64	26	310	50	20
2. Sig *r*	1,163	125	41	783	85	28	617	68	22
3. *r* dif	2,339	263	96	1,573	177	66	1,240	140	52
4. *P* = .5	1,165	127	44	783	85	30	616	67	23
5. *P* dif	584	93	36	392	63	25	309	49	19
6. χ^2									
1*df*	1,168	130	38	785	87	26	618	69	25
2*df*	1,388	154	56	964	107	39	771	86	31
3*df*	1,546	172	62	1,090	121	44	880	98	35
4*df*	1,675	186	67	1,194	133	48	968	108	39
5*df*	1,787	199	71	1,293	143	51	1,045	116	42
6*df*	1,887	210	75	1,362	151	54	1,113	124	45
7. ANOVA									
2*g*[a]	586	95	38	393	64	26	310	50	20
3*g*[a]	464	76	30	322	52	21	258	41	17
4*g*[a]	388	63	25	274	45	18	221	36	15
5*g*[a]	336	55	22	240	39	16	193	32	13
6*g*[a]	299	49	20	215	35	14	174	28	12
7*g*[a]	271	44	18	195	32	13	159	26	11
8. Mult *R*									
2*k*[b]	698	97	45	481	67	30			
3*k*[b]	780	108	50	547	76	34			
4*k*[b]	841	118	55	599	84	38			
5*k*[b]	901	126	59	645	91	42			
6*k*[b]	953	134	63	686	97	45			
7*k*[b]	998	141	66	726	102	48			
8*k*[b]	1,039	147	69	757	107	50			

Note. ES = population effect size, Sm = small, Med = medium, Lg = large, diff = difference, ANOVA = analysis of variance. Tests numbered as in Table 1.
[a] Number of groups. [b] Number of independent variables.

either *r* is smaller then .30 or the investigator has been the victim of the .20 (β) risk of making a Type II error.

Examples

The necessary *N* for power of .80 for the following examples are found in Table 2.

1. To detect a medium difference between two independent sample means (d = .50 in Table 1) at α = .05 requires N = 64 in each group. (A d of .50 is equivalent to a point-biserial correlation of .243; see SPABS, pp. 22–24.)

2. For a significance test of a sample *r* at α = .01, when the population *r* is large (.50 in Table 2), a sample size = 41 is required. At α = .05, the necessary sample size = 28.

3. To detect a medium-sized difference between two population *r*s (q = .30 in Table 1) at α = .05 requires N = 177 in each group. (The following pairs of *r*s yield q = .30: .00, .29; .20, .46; .40, .62; .60, .76; .80, .89; .90, .94; see SPABS, pp. 113–116.)

4. The sign test tests the H_0 that .50 of a population of paired differences are positive. If the population proportion's departure from .50 is medium (q = .15 in Table 1), at α = .10, the necessary N = 67; at α = .05, it is 85.

5. To detect a small difference between two independent population proportions (h = .20 in Table 1) at α = .05 requires N = 392 cases in each group. (The following pairs of *P*s yield approximate values of h = .20: .05, .10; .20, .29; .40, .50; .60, .70; .80, .87; .90, .95; see SPABS, p. 184f.)

6. A 3 × 4 contingency table has 6 degrees of freedom. To detect a medium degree of association in the population (w = .30 in Table 1) at α = .05 requires N = 151. (w = .30 corresponds to a contingency coefficient of .287, and for 6 degrees of freedom, a Cramèr ϕ of .212; see SPABS, pp. 220–227).

7. A psychologist considers alternate research plans involving comparisons of the means of either three or four groups in both of which she believes that the ES is medium (f = .25 in Table 1). She finds that at α = .05, the necessary sample size per group is 52 cases for the three-group plan and 45 cases for the four-group plan, thus, total sample sizes of 156 and 180. (When f = .25, the proportion of variance accounted for by group membership is .0588; see SPABS, pp. 280–284.)

8. A psychologist plans a research in which he will do a multiple regression/correlation analysis and perform all the significance tests at α = .01. For the *F* test of the multiple R^2, he expects a medium ES, that is, f^2 = .15 (from Table 1). He has a candidate set of eight independent variables for which Table 2 indicates that the required sample size is 147, which exceeds his resources. However, from his knowledge of the research area, he believes that the information in the eight variables can be

effectively summarized in three. For three variables, the necessary sample size is only 108. (Given the relationship between f^2 and R^2, the values for small, medium, and large R^2 are respectively .0196, .1304, and .2592, and for R, .14, .36, and .51; see SPABS, pp. 410–414.)

References

Cohen, J. (1962). The statistical power of abnormal–social psychological research: A review. *Journal of Abnormal and Social Psychology, 65,* 145–153.

Cohen, J. (1965). Some statistical issues in psychological research. In B. B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95–121). New York: McGraw-Hill.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences.* San Diego, CA: Academic Press.

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

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304–1312.

Goldstein, R. (1989). Power and sample size via MS/PC-DOS computers. *American Statistician, 43,* 253–260.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika, 20A,* 175–240, 263–294.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Transactions of the Royal Society of London Series A, 231,* 289–337.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309–316.

Received February 1, 1991
Revision received April 26, 1991
Accepted May 2, 1991 ■