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 rs, (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 accidental of the historical precedence of Fi.
sherian theory, its hybridization with the contradictory Ney-
man-Pearson theory, and the apparent completeness of Fisher-
ian 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 

An associate editor of this journal suggests another reason: 
Researchers find too complicated, or do not have at hand, ei-
ther my book or other reference material for power analysis. He 
suggests that a short rule-of-thumb treatment of necessary sam-
ple 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 unhesitat-
ingly 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 ex-
amples 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 \( N \), signifi-
cance criterion \( \alpha \), 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 \( \alpha \), \( N \), and \( ES \). For research 
planning, however, it is most useful to determine the \( N \) necessary to 
have a specified power for given \( \alpha \) and \( ES \); this article addresses 
this use.

**The Significance Criterion, \( \alpha \)**

The risk of mistakenly rejecting the null hypothesis \( H_0 \) and thus of 
committing a Type I error, \( \alpha \), 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 sev-
eral \( H_0 \)s, 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 \( \alpha \) 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 de-
sired, as, for example, in exploratory studies. For unreconstructed one 
tailers (see Cohen, 1965), the tabled sample sizes provide close approxi-
mations for one-sided tests at .10 (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 probabil-
ity, given the population \( ES \), \( \alpha \), 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 \( ES \), \( \alpha \), and \( N \), its probability of 
occurring is \( \beta \). Power is \( 1 - \beta \), 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 
\( \beta \)x 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 \( N \) neces-
sary to attain the desired power for the specified \( \alpha \) and hypothesized 
\( ES \). \( N \) increases with an increase in the power desired, a decrease in the 
\( ES \), and a decrease in \( \alpha \). 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). How-
ever, 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 
counterposed against \( H_0 \). The degree to which \( H_1 \) is false is indexed by the discrep-
ancy 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 indepen-
dent means, correlation coefficients, and proportions, the \( H_0 \) that 
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 conven-
tions have been fixed since the 1977 edition of SPABS and have come 
together in general use. Table 1 contains these values for the tests considered 
here.

In the present treatment, the \( H_1 \)s are the \( ES \)s that operationally de-
fine 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 
and large \( ES \)s 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

<table>
<thead>
<tr>
<th>Test</th>
<th>ES index</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. $m_A$ vs. $m_B$ for independent means</td>
<td>$d = \frac{m_A - m_B}{\sigma}$</td>
<td>0.20 0.50 0.80</td>
</tr>
<tr>
<td>2. Significance of product-moment $r$</td>
<td>$r$</td>
<td>0.10 0.30 0.50</td>
</tr>
<tr>
<td>3. $r_A$ vs. $r_B$ for independent $r$s</td>
<td>$q = z_A - z_B$ where $z =$ Fisher's $z$</td>
<td>0.10 0.30 0.50</td>
</tr>
<tr>
<td>4. $P = .5$ and the sign test</td>
<td>$g = P - .50$</td>
<td>0.05 0.15 0.25</td>
</tr>
<tr>
<td>5. $P_A$ vs. $P_B$ for independent proportions</td>
<td>$h = \phi_A - \phi_B$ where $\phi =$ arcsine transformation</td>
<td>0.20 0.50 0.80</td>
</tr>
<tr>
<td>6. Chi-square for goodness of fit and contingency</td>
<td>$w = \sqrt{\frac{\sum (P_{ui} - P_{oi})^2}{P_{oi}}}$</td>
<td>0.10 0.30 0.50</td>
</tr>
<tr>
<td>7. One-way analysis of variance</td>
<td>$f = \frac{\sigma_m}{\sigma}$</td>
<td>0.10 0.25 0.40</td>
</tr>
<tr>
<td>8. Multiple and multiple partial correlation</td>
<td>$f^2 = \frac{R^2}{1 - R^2}$</td>
<td>0.02 0.15 0.35</td>
</tr>
</tbody>
</table>

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 $H_0$) are $d = .20, .50,$ and .80. Thus, an operationally defined medium difference between means is half a standard deviation, 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 $r$s, accomplished as a normal curve test through the Fisher $z$ transformation of $r$ (tabulated 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 $\phi$ (tabulated 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 $P_A$ and $P_B$ 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 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 $a = .05$, 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

<table>
<thead>
<tr>
<th>Test</th>
<th>α = .01</th>
<th>α = .05</th>
<th>α = .10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sm</td>
<td>Med</td>
<td>Lg</td>
</tr>
<tr>
<td>1. Mean diff</td>
<td>586</td>
<td>95</td>
<td>38</td>
</tr>
<tr>
<td>2. Sig r</td>
<td>1,163</td>
<td>125</td>
<td>41</td>
</tr>
<tr>
<td>3. r diff</td>
<td>2,339</td>
<td>263</td>
<td>96</td>
</tr>
<tr>
<td>4. P = .05</td>
<td>1,165</td>
<td>127</td>
<td>44</td>
</tr>
<tr>
<td>5. P diff</td>
<td>584</td>
<td>93</td>
<td>36</td>
</tr>
<tr>
<td>6. χ²</td>
<td>1,168</td>
<td>130</td>
<td>38</td>
</tr>
<tr>
<td>1 df</td>
<td>1,388</td>
<td>154</td>
<td>56</td>
</tr>
<tr>
<td>2 df</td>
<td>1,546</td>
<td>172</td>
<td>62</td>
</tr>
<tr>
<td>3 df</td>
<td>1,675</td>
<td>186</td>
<td>67</td>
</tr>
<tr>
<td>5 df</td>
<td>1,787</td>
<td>199</td>
<td>71</td>
</tr>
<tr>
<td>6 df</td>
<td>1,887</td>
<td>210</td>
<td>75</td>
</tr>
<tr>
<td>7. ANOVA</td>
<td>2g</td>
<td>586</td>
<td>95</td>
</tr>
<tr>
<td>3g</td>
<td>464</td>
<td>76</td>
<td>30</td>
</tr>
<tr>
<td>4g</td>
<td>388</td>
<td>63</td>
<td>25</td>
</tr>
<tr>
<td>5g</td>
<td>330</td>
<td>55</td>
<td>22</td>
</tr>
<tr>
<td>6g</td>
<td>299</td>
<td>49</td>
<td>20</td>
</tr>
<tr>
<td>7g</td>
<td>271</td>
<td>44</td>
<td>18</td>
</tr>
<tr>
<td>8. Multi R</td>
<td>2k</td>
<td>698</td>
<td>97</td>
</tr>
<tr>
<td>3k</td>
<td>780</td>
<td>108</td>
<td>50</td>
</tr>
<tr>
<td>4k</td>
<td>841</td>
<td>118</td>
<td>55</td>
</tr>
<tr>
<td>5k</td>
<td>901</td>
<td>126</td>
<td>59</td>
</tr>
<tr>
<td>6k</td>
<td>953</td>
<td>134</td>
<td>63</td>
</tr>
<tr>
<td>7k</td>
<td>998</td>
<td>141</td>
<td>66</td>
</tr>
<tr>
<td>8k</td>
<td>1,039</td>
<td>147</td>
<td>69</td>
</tr>
</tbody>
</table>

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.

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 SPAB, 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 rs (r = .30 in Table 2) at α = .05 requires N = 177 in each group. (The following pairs of rs yield r = .30: .00, .29, .20, .46; .40, .62; .60, .76; .80, .89; .90, .94; see SPAB, pp. 113-116.)

4. The sign test tests the H0 that .50 of a population of paired differences are positive. If the population proportion's departure from .50 is medium (r = .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 Ps yield approximate values of h = .20: .05, .10, .20, .29; .40, .50, .60, .70; .80, .87; .90, .95; see SPAB, p. 184f)

6. A 3 × 4 contingency table has 6 degrees of freedom. To detect a medium degree of association in the population (w = .50 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 SPAB, 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 SPAB, 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

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

---

**Low Publication Prices for APA Members and Affiliates**

**Keeping You Up-to-Date:** All APA members (Fellows; Members; Associates, and Student Affiliates) receive—as part of their annual dues—subscriptions to the *American Psychologist* and *APA Monitor*.

High School Teacher and International Affiliates receive subscriptions to the *APA Monitor*, and they can subscribe to the *American Psychologist* at a significantly reduced rate.

In addition, all members and affiliates are eligible for savings of up to 60% (plus a journal credit) on all other APA journals, as well as significant discounts on subscriptions from cooperating societies and publishers (e.g., the American Association for Counseling and Development, Academic Press, and Human Sciences Press).

**Essential Resources:** APA members and affiliates receive special rates for purchases of APA books, including the *Publication Manual of the APA*, the *Master Lectures*, and *Journals in Psychology: A Resource Listing for Authors*.

**Other Benefits of Membership:** Membership in APA also provides eligibility for low-cost insurance plans covering life, income protection, office overhead, accident protection, health care, hospital indemnity, professional liability, research/academic professional liability, student/school liability, and student health.

**For more information,** write to American Psychological Association, Membership Services, 750 First Street, NE, Washington, DC 20002-4242, USA