TESTING RESEARCH HYPOTHESES WITH THE GENERAL LINEAR MODEL

Power Analysis

Keith McNeil,
Isadore Newman,
and Francis J. Kelly

1996

Southern Illinois University Press
Carbondale and Edwardsville

INTRODUCTION TO POWER

Power is the ability of the statistical analysis to find significance if in fact significance is there. There are various ways to increase power:

1. Increase the number of subjects.
2. Change alpha from, say, .05 to .10.
3. Increase the difference between the Statistical Hypothesis and reality (e.g., if $\mu = 50$, then a Research Hypothesis of $\mu > 45$ is less powerful than a Research Hypothesis of $\mu > 40$).
4. Make the error term smaller:
   a. by using measures that are more reliable and more valid;
   b. by including individual differences if they are in the design;
   c. by blocking on known, relevant variables;
   d. by including nonlinear and interaction predictor variables if thought relevant; and
   e. by covarying initial differences between groups.

Some of these concepts have already been discussed; others will be discussed in the next chapter. The material that follows has been adapted from a paper by Newman and Benz (1980), which goes into more detail, particularly on the calculation of the required number of subjects and the combination of study results.

What follows is a discussion of how to use Cohen's (1977) power analysis tables (see Appendix G). There are four parameters that one must be aware of when conducting a power analysis:

alpha	α
sample size	N
effect size	f^2
power	power

If one knows any three of the four, one can solve for the fourth.

Alpha

Alpha (α) is the probability of making a Type I error. It is under the control of the researcher and is generally set at .05, .01, or .001.

Effect Size

Effect size can be thought of as how far apart the means of two groups are in terms of standard deviation units (e.g., 1.2 standard deviations, .50 standard deviations). Another way of looking at it is in terms of the proportion of variance accounted for. In correlational analyses, the r^2 (if using Pearson correlation)—or R^2 (if using the GLM)—would provide this information; and in ANOVA designs, the ω^2 would provide the information (ω^2 being the symbol used in ANOVA to represent the proportion of variance accounted for).

Cohen (1977) uses f^2 to represent effect size and arbitrarily defines three effect sizes: large (>.35), medium (.15 to .35), and small (<.15). Effect size in reality is set depending on how well the researcher knows her field of research and what she is

looking for. Effect sizes that are considered large in one instance may be considered small in another.

Power

Power is the probability of detecting a population fact when in fact that population fact exists. Because the probability of not detecting a population fact is defined as the probability of making a Type II error, power is then one minus the probability of making a Type II error. For example, if the power of a test is .76, this means that 76 times out of 100 the statistical procedures (given the researcher's choice of *N*, alpha, and effect size) will be capable of detecting the population fact if it exists.

Sample Size

N is the total number of subjects used in the study.

Calculating Power

To use Appendixes G-1 through G-6, two variables need to be determined. First, the df_n needs to be computed. Remember, df_n is equal to (m1 - m2), the pieces of information in the Full Model minus the pieces of information in the Restricted Model. Cohen labels this value as U. Second, the variable L needs to be calculated. The variable L has no direct meaning but it is used as one of the entries for Appendixes G-1 through G-6.

$$L = f^2 * V$$

where:

f^2 = effect size; and
V = df_d = (*N* - m1) = (*N* minus pieces of information in Full Model).

There is a separate table for each of the three different alpha values .01, .05, and .10. Assume we have 100 subjects (*N* = 100) and we want to be able to detect a medium-size effect (f^2 = .15). Assume also that we have 10 linearly independent variables (including the unit vector) in the Full Model. We are interested in asking the following question: "Do these 10 variables account for variance in the criterion, over and above no information?" Let us assume our alpha level is set at .01. We now can determine power.

$f^2 = .15$

V = (*N* - m1) = (100 - 10) = 90

$L = f^2 * V$
$L = (.15) * (90)$
$L = 13.5$

U = (m1 - m2) = (10 - 1) = 9

alpha = .01

Since alpha is .01, we would use Appendix G-1. We enter the table at a U of 9. We look for an L value of 13.5, which falls between L values of 12.00 and 14.00, requiring an interpolation to obtain an estimated power of 49.5.

If we are interested in doing this problem at an alpha of .05, we would use Appendix G-2. Looking at U = 9, L is between 12.00 and 14.00, and we have an estimated power of 72. We can see that as alpha becomes less stringent (.01 to .05 to .10), the power increases from 49% to 72% to 81%.

Solving for *N*

Given the same research question, we can determine the *N* size for a given effect size. Solving the previous power equation for *N* yields: $N = (L / f^2) + m1$. For this problem we have (a) alpha = .05; (b) m1 = 10; (c) $f^2 = .02$; (d) U = 9; and (e) power = .80. (Cohen recommends a power of .80 if no other information is given, which is a comparable rationale to setting alpha equal to .05.) To determine the L size for an alpha of .05, we use Appendix G-5. We enter the table for a given power and a particular U. Since power is set at .80 and U is 9, our L is 15.65. Using the above formula, we solve for *N*:

$$N = (15.56 / .02) + 10$$
$$N = 788$$

So, 788 subjects would be required to detect a small effect size (.02) with an alpha of .05 with this question (using 10 predictor variables). Whenever one solves for *N* and the value has a decimal, one always rounds upwards, so if *N* had been, say, 792.5, *N* would be equal to 793. If an alpha of .01 were desired, the number of subjects required would be 1,082.

When deciding on the desired effect size, one may adopt the large, medium, or small effect sizes arbitrarily identified by Cohen, or one may look at the effect size reported in the literature for that topic. One could randomly sample 10 articles, calculate the effect sizes, take the average, and use that as a yardstick, a starting point. However, most research contains a test of significance (such as a *t* test, *z* test, *F* test, or chi-square test) instead of effect size. How does one change these different tests of significance to effect sizes? The astute GLM reader is aware that these analyses could have initially been accomplished with the GLM; fortunately, formulae exist to change these reported test statistics to effect sizes.

A study reporting a *t* test also may report the amount of variance accounted for with a point biserial correlation coefficient; a study employing an *F* test may report an eta coefficient; and one employing a chi-square test may report a phi or contingency coefficient (see Table 8.1). These all represent the proportion of variance accounted for and can be represented as r_m.

All of these tests of significance can be transferred into the effect size, f^2. This is done through the intermediary components of S and Q^2. Once S and Q^2 have been computed, the value r_m can be computed as:

$$r_m = Q^2 / (Q^2 + S)$$

The square of r_m, r_m^2, is the proportion of variance accounted for. The effect size in the literature can then be calculated by the following formula:

Needed to Calculate r_m

Q	Q^2	S	Correlational Analysis
t	NA	*df*	point biserial
z	NA	*N*	point biserial
NA	$F * (df_n)$	df_d	eta
NA	χ^2	*N*	phi or contingency
NA	$F * (df_n)$	df_d	Multiple R

$$f^2 = \frac{r_m^2}{1 - r_m^2}$$

What is interesting is that when r_m is small, the effect size and r are almost identical. But when r_m is large, effect size approaches infinity as r_m approaches 1.

Suppose that a search of the literature produced three studies with these results:

Study #1: $t = 4$, $df = 84$;

Study #2: $F = 3$, $df_n = 2$, $df_d = 94$;

Study #3: $\chi^2 = 4$, $N = 96$.

The following shows how to convert the results from these three studies to r_m.

Study #1 (t test): t to r_m (where: $t = 4$ and $df = 84$)

$$r_m = \frac{Q^2}{Q^2 + S} = \frac{4^2}{4^2 + 84} = \frac{16}{100} = .16$$

Study #2 (ANOVA): F to r_m (where: $F = 3$, $df_n = 2$, and $df_d = 94$)

Please note that since the F in Table 8.1 is under Q^2, we do not have to square the F as we did for the t in Study #1.

$$r_m = \frac{Q^2}{Q^2 + S} = \frac{(3 * 2)}{(3 * 2) + 94} = \frac{6}{100} = .06$$

Study #3 (chi-square): χ^2 to r_m (where: $\chi^2 = 4$, and $N = 96$)

$$r_m = \frac{Q^2}{Q^2 + S} = \frac{4}{4 + 96} = \frac{4}{100} = .04$$

Each of these r_m can then be transformed to an effect size through the following formula:

$$f^2 = \frac{r_m^2}{1 - r_m^2}$$

	r_m	r_m^2	f^2
Study #1 (*t* test):	.16	.0256	.03
Study #2 (ANOVA):	.06	.0036	.004
Study #3 (chi-square):	.04	.0016	.0016

The average effect size for the above three studies could be calculated to provide guidance to the researcher on the choice of effect size from three apparently disparate types of statistical results. It should be clear, though, that the effect size for these three studies all fall within Cohen's "small" definition. Here we have evidence of what has been alluded to in previous chapters. One can obtain statistical significance, but there may not be any practical significance. Finally, one also could use the effect-size equation to calculate the r_m^2 from a published effect size:

$$f^2 = \frac{r_m^2}{1 - r_m^2}$$

$$f^2 * (1 - r_m^2) = r_m^2$$

$$f^2 - (f^2 * r_m^2) = r_m^2$$

$$f^2 = r_m^2 + (f^2 * r_m^2)$$

$$f^2 = r_m^2 * (1 + f^2)$$

$$\frac{f^2}{1 + f^2} = r_m^2$$

ANCOVA to Obtain Increased Power

Researchers are urged to use ANCOVA even when random assignment is made and even when the group means of the covariate are identical. Mueller (1990) discusses this use of ANCOVA and states concerns for using the technique when groups have not been randomly assigned. We, in this text, take the position that the statistical technique is unaware of whether the subjects have been randomly assigned. Making a covariance adjustment is better than not making one. Ultimately, research must be conducted with manipulated *and* intact groups. The ANCOVA can provide insight until that time.

In cases where pretest ability is known, the inclusion of these data in an over and above analysis will usually provide a better estimate of within-group variance ($\hat{v}_w$). Usually people who score high on the pretest will score relatively high on the posttest, and those who score low on the pretest will usually score fairly low on the posttest. The correlation between pretest and posttest is often greater than zero. Therefore, the R^2 of a model containing knowledge of both treatment and pretest scores will be larger than an R^2 of a model using only knowledge of treatment. The over and above test still tests the unique contribution of the independent variable (e.g., treatment), but the proportional estimate of the population variance within $[(1 - R_f^2)/(N - m1)]$ will be smaller at the expense of having only one degree of freedom (due to including the covariate) and at the cost of collecting the covariate score. Essentially, the reasoning is, "Why throw away knowledge regarding the sample when one has it?" If the task of the researcher is to attempt to account for as much of the variance as possible (R^2 as close to 1.0 as possible), then one might go further and recommend that the researcher include many covariates that account for nonrandom criterion variance. Chapter 1 was written with this viewpoint.

A word of caution is in order here. The ideal covariate is one that is *not correlated* with the overlap between the criterion and the predictor(s) but *is correlated* with the error in those predictions. If the covariate is correlated with the predictor part of the overlap between predictors and criterion, then the predictors' effects are confounded with the covariate. In this case, the predictor effects will be attributed to the covariate, and what might have originally been an effective predictor may now be wiped out. Such are the trials and temptations of research in the behavioral sciences.

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

Isadore Newman

Table 1

All tests of significance are tests of relationships

Tests of Significance	Measures of Relationships
t test	r_{pb}; that is $t^2 = \frac{r_{pb}^2\, df}{1 - r_{pb}^2}$
z test	r_{pb} (point biserial)
F test	η^2 (eta); could also be measured by R^2
χ^2 test	ϕ^2 which is $\frac{\chi^2}{N} = \phi^2$ (phi coefficient)
	when df = 1 or greater than 1 (contingency coefficient) $C^2 = \frac{\chi^2}{\chi^2 + N}$ Cramer. $V = \sqrt{\frac{\chi^2}{N(n-1)}}$ # of Rows or Columns, whichever is smaller