pression, *Journal of Clinical Psychiatry, 51,* 61–69 (1990).

11. L.R. Baxter, Jr., J.M. Schwartz, B.H. Guze, J.C. Mazziotta, M.P. Szuba, K. Bergman, A. Alazraki, C.E. Selin, H.K. Freng, P. Munford, and M.E. Phelps, *Obsessive-compulsive disorder vs. Tourette's disorder: Differential function in subdivisions of the neostriatum,* paper presented at the annual meeting of the American College of Neuropsychopharmacology, San Juan, Puerto Rico (December 1991).

12. E.M. Reiman, M.E. Raichle, F.K. Butler, P. Herscovitch, and E. Robins, A focal brain abnormality in panic disorder, a severe form of anxiety, *Nature, 310,* 683–685 (1984); E.M. Reiman, M.E. Raichle, E. Robins, F.K. Butler, P. Herscovitch, P. Fox, and J. Perlmutter, The application of positron emission tomography to the study of panic disorder, *American Journal of Psychiatry, 143,* 469–477 (1986); T.E. Nordahl, W.E. Semple, M. Gross, T.A. Mellman, M.B. Stein, P. Goyer, A.C. King, T.W. Uhde, and R.M. Cohen, Cerebral glucose metabolic differences in patients with panic disorder, *Neuropsychopharmacology, 3,* 261–272 (1990).

Statistical Power Analysis

Jacob Cohen

The power of a statistical test of a null hypothesis (H_0) is the probability that the H_0 will be rejected when it is false, that is, the probability of obtaining a statistically significant result. Statistical power depends on the significance criterion (α), the sample size (N), and the population effect size (ES).

The importance of power analysis arises from the fact that most empirical research in the social and behavioral sciences proceeds by formulating and testing H_0s that the investigators hope to reject as a means of establishing facts about the phenomena under study.

A typical H_0 is that a population product-moment correlation, r, is zero, to be tested at the two-sided (α_2 =) .05 level. When this H_0 is tested on a sample of N cases randomly drawn from a population in which r indeed does equal zero, researchers risk mistakenly rejecting the H_0 when it is true, a *Type I* error, whose rate (.05) is controlled by the α criterion. They also risk mistakenly accepting the H_0 as tenable when it is false, a *Type II* error, whose probability is called β. Power is thus $1 - \beta$, the probability of *not* accepting the H_0 when it is false, that is, the probability of successfully rejecting the H_0.

The outcome of a statistical test depends on the degree to which the H_0 is false, that is, on the magnitude of the population ES, which in this case is the absolute size of the population r—the larger the r, the greater the likelihood that the H_0 will be rejected. It is also true that the outcome depends on N, a larger sample being more likely to result in rejection of a false H_0 than a smaller one. Thus, at α_2 = .05, for example, if the population r is .30, when N is 40, the power of the standard t test of a sample r turns out to equal .48, whereas when N is 80, power is .78. If the population r is .40, when N is 40, power is .74, but when N is 80, power is .96. Finally, the test outcome depends also on α, the risk of a Type I error. A smaller and therefore more stringent α criterion, say, α_2 = .01, for any given population r and N, would result in smaller power. For example, with population r = .30 and N = 80, while power at α_2 = .05 is .78, power at α_2 = .01 is only .56.[1] Note also that at any given value of α, a two-sided test is more stringent than a one-sided test.

Statistical power analysis exploits the mathematical relationship among these four variables in statistical inference: power, α, N, and ES. The relationship is such that when any three of them are fixed, the fourth is determined. Two forms of power analysis are most useful: One is the determination of the N that is necessary to attain a specified degree of power to detect as significant (at specified α) a hypothesized ES. This form of power analysis is used in research planning. The second is the determination of power to detect a hypothesized ES (for specified N and α), the form used in meta-analytic power reviews of research areas or journals.

Jacob Cohen, Professor of Psychology at New York University, is the author of *Statistical Power Analysis for the Behavioral Sciences* (2nd ed., 1988) and co-author with Patricia Cohen of *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (2nd ed., 1983), both published by Lawrence Erlbaum Associates. Address correspondence to Jacob Cohen, Department of Psychology, New York University, 6 Washington Place, 5th Floor, New York, NY 10003.

EFFECT SIZE

I noted that in testing a sample r, the ES is simply the population r. More generally, in the Neyman-Pearson system of statistical induction,[2] whence the concept of power is derived, the ES is the discrepancy between the null hypothesis, H_0, and the alternate hypothesis of interest, H_1. For testing a sample r, the H_0 is that the population r is zero, and the H_1 posits a specific nonzero value, for example, .30. Thus, the ES in this example is simply the difference: .30 − .00. Every statistical test has its own ES index, a continuous value that runs from zero,

when the H_0 is true. Each ES index is a pure (i.e., scale-free) value that measures, in terms appropriate to it, the discrepancy between the H_0 and the H_1.

For example, the ES index for the difference between independent means in the classical *t* test is *d*, the difference between the population means standardized by dividing this difference by the common within-population standard deviation. (The difference is absolute for two-sided tests and is either positive or negative for one-sided tests.) The standardization results in a scale-free measure: $d = .25$ implies a quarter of a standard deviation difference between the population means, free of the units of measure of the variable in question, whether they are inches, centimeters, or points scored on a psychological test.

As another example, for testing the departure of a population proportion (*P*) from .50, the ES index is $g = P - .50$. If an investigator believes that there is a sex difference in the incidence of dyslexia such that boys are at different risk from girls, in a sample of dyslexic children, she would posit as the H_0 that half the sample are of one sex, and as the H_1 that a specified different proportion, say, .60, are of the other. The ES index would then be $g = .60 - .50 = .10$. Still another example is the analysis of variance test that a set of population means are all equal. The ES index for this test, *f*, is the standard deviation of these means divided by the common within-population standard deviation of the observations.[1]

Investigators in the social sciences find specifying the ES the most difficult aspect of power analysis. This is at least partly due to the relatively low level of consciousness about magnitudes in those disciplines. The conquest of psychological science by Fisherian null hypothesis testing (where the alternative to the H_0 is simply its negation, so that no H_1 is specified) has had the unfortunate effect of emphasizing the magnitudes of *p* values from significance tests rather than the magnitudes of the psychological *phenomena* under study.[3] A salutary side effect of the study of power analysis is its emphasis on ES. Neither power nor sample size can be determined in the absence of the investigator's readiness to consider just how wrong the null hypothesis is likely to be (i.e., the ES). The decision as to what population ES to posit arises from the investigator's knowledge of the field—the sample ESs found in previous investigations with similar variables, the results of pilot studies (though not reliable when based on small samples), and his or her educated intuition.

Because the ES indices are not generally familiar, I have proposed as conventions, or operational definitions, "small," "medium," and "large" values of each ES index to provide the user with some sense of its scale.[1] It was my intent that medium ES represent an effect of a size likely to be apparent to the naked eye of a careful observer, that small ES be noticeably smaller yet not trivial, and that large ES be the same distance above medium as small is below it. I also made an effort to make these conventions comparable across different statistical tests.

For example, for the test that $r = 0$, small, medium, and large ESs are, respectively, the population *r*s .10, .30, and .50. For the test that two population means are equal, the ESs, in the same order, are $d = .20$, .50, and .80. The .20 ES is exemplified by the mean IQ difference between twins and nontwins (the latter being larger), the .50 ES by the mean IQ difference between clerical and semiskilled workers, and the .80 ES by the mean IQ difference between Ph.D.s and college freshmen. In the analysis of variance test of the H_0 that *g* populations have equal means, the index *f*s (the standardized standard deviation of the means) are, respectively, .10, .25, and .40 for small, medium, and large ESs.[1]

Another means of facilitating the understanding of the various ES indices is by transforming them into other measures. For example, many of the ES indices (e.g., *d*, *f*, and the ES indices for the difference between proportions and for the degree of association in contingency tables of frequencies) may be translated into correlation coefficients or their squares, which may then be interpreted as proportions of variance. As another example, *d* may be expressed as various proportions of (non)overlap between normal distributions.[1]

α, THE SIGNIFICANCE CRITERION

The probability of mistakenly rejecting the H_0, α, represents a research policy—the maximum risk one is prepared to take of making this error. It has become conventional that unless otherwise stated, this risk is set at .05. Smaller and thus more stringent values may be used, for example, when several H_0s are to be tested in order to minimize the risk of making *any* Type I errors in investigation (the *experimentwise* risk). Larger values may be used in exploratory studies. Also, for tests whose ESs may be either positive or negative, α may be defined as two-sided or one-sided. The latter has more power than the former when the sample effect is in the direction posited, but has zero power when the effect is in the opposite direction because the one-sided test logically precludes a contrary finding.

DETERMINING SAMPLE SIZE

In planning research, deciding the sample sizes is crucial. Because

research costs are at least approximately linear in the number of subjects, cost-effectiveness demands that this decision be appropriate. When asked in connection with a particular investigation what α and power are desired, a neophyte researcher might suggest $\alpha_2 = .01$ and some very large value for power, say, .99. Power analysis quickly determines that these specifications necessitate a sample size that is likely beyond the available resources. For example, for a test of the difference between means, if a medium ES ($d = .5$) exists in the population, these specifications require 194 cases in each of the two samples. Similarly, they require that if population $r = .30$, a test of the significance of a sample r have 254 cases. For $\alpha_2 = .05$ and .99 power, the N requirements are, respectively, 148 and 195.

To determine the necessary sample size, one needs to posit the α, ES, and desired power. I have proposed as a convention that in the absence of any other basis for setting the value for desired power, .80 be used.[1] In scientific research, it is typically more serious to make a false positive claim (Type I error) than a false negative one (Type II error). Because the implicit convention for significance is $\alpha = .05$, the use of the .80 convention for desired power (hence, $\beta = .20$) makes the Type II error 4 times as likely as the Type I error, an arbitrary but reasonable reflection of their relative importance.[4]

A useful aid in determining the necessary sample size is a sample size planning table. To prepare such a table, the investigator selects values or ranges of values for α, ES, and power and then determines the N for each combination. This table provides the basis for a judicious choice of specifications or leads to the useful discovery that the research as conceived is not viable.[3]

Recall the investigator pursuing the question of a sex difference in the incidence of dyslexia. If in a population of dyslexic children half are boys, there is no sex difference, so H_0 is $P = .50$. Departure from .50 would render H_0 false. The ES index for this test is $g = P - .50$, the departure of the proportion from one half. If the investigator's resources are such that she could obtain an N of 90 to 100, and her expectation is a value of g in the range .10 to ⅙, she might compile the sample size planning table shown in Table 1 by looking up various combinations of α_2 and g that would result in Ns within the desired range and noting the resulting power. From this table, she could choose a set of specifications.

DETERMINING POWER

There is a useful role for power analysis in assessing completed research, particularly research in which nonsignificant results were obtained. Given the N employed and α, one needs only to posit the population ES to determine power. The sample ES found, or one or more ES values posited by the assessor, may serve this purpose. It is a common finding that power was poor for plausible ESs, usually because of small N.

In 1962, I reviewed the articles in the 1960 volume of the *Journal of Abnormal and Social Psychology* from the perspective of power.[5] I determined power for each statistical test in each article using the N employed at $\alpha_2 = .01$, .05, and .10 for the conventional definitions of small, medium, and large ES. I found, for example, that the median power to detect a medium ES at $\alpha_2 = .05$ was .46. The many power surveys done in the biosocial sciences since that time have had similar results. For example, a similar review by Sedlmeier and Gigerenzer of the 1984 *Journal of Abnormal Psychology*[6] found the median power under the same conditions to be a little worse (.44)—and it was lower still (.37) when an experimentwise α criterion was employed. Even worse was the finding that in 11% of the studies, the H_0 was taken as the research hypothesis and nonsignificance taken as confirmation: The median power of these studies to detect a medium ES at $\alpha_2 = .05$ was .25!

Table 1. *A sample size planning table*

α_2	g	Power	N
.01	1/6	.75	92
.02	.15	.75	98
.02	1/6	.85	98
.05	.10	.50	96
.05	.15	.85	97
.10	.10	.60	90
.10	.15	.90	91
.10	1/6	.95	92
.20	.15	.95	90

CONCLUSION

There has been no disagreement among research methodologists about the desirability of power analysis in research planning and assessment, yet progress in application of this method over the last quarter century has been slow. There have, however, been some rays of hope in the past few years. The popularity of meta-analysis has served to emphasize the size of effects and by thus raising the consciousness of behavioral scientists has promoted the cause of power analysis.[3] More directly, both graduate and undergraduate statistics textbooks have begun to feature chapter-length treatments of power analysis.[7] Finally, in addition to the reference works already noted,[1,4] there are available computer programs for power analysis and sample size determination.[8]

Acknowledgments—I am, as always, grateful to Patricia Cohen for her helpful comments.

Notes

1. J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. (Erlbaum, Hillsdale, NJ, 1988). This is the source of the system of power analysis described here; the power values and sample sizes of the illustrations derive from this book's tables.

2. J. Neyman and E.S. Pearson, On the use and interpretation of certain test criteria for purposes of statistical inference, *Biometrika, 20A*, 175–240, 263–294 (1928); J. Neyman and E.S. Pearson, On the problem of the most efficient tests of statistical hypotheses, *Transactions of the Royal Society of London Series A, 231*, 289–337 (1933).

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

4. For an article-length treatment of sample size determination using the .80 convention and α = .01, .05, and .10, see J. Cohen, A power primer, *Psychological Bulletin* (in press). A useful alternative treatment is offered in H.C. Kraemer and S. Thiemann, *How Many Subjects? Statistical Power Analysis in Research* (Sage, Newbury Park, CA, 1987).

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

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

7. R. Rosenthal and R.L. Rosnow, *Essentials of Behavioral Research: Methods and Data Analysis*, 2nd ed. (McGraw Hill, New York, 1991); J. Welkowitz, R.B. Ewen, and J. Cohen, *Introductory Statistics*, 4th ed. (Harcourt Brace Jovanovich, San Diego, 1991).

8. M. Borenstein and J. Cohen, *Statistical Power Analysis: A Computer Program* (Erlbaum, Hillsdale, NJ, 1988); J. Hintze, *Power Analysis and Sample Size* (NCSS, Kaysville, UT, 1991). Some 13 programs are reviewed in R. Goldstein, Power and sample size via MS/PC-DOS computers, *American Statistician, 43*, 253–260 (1989).

Why Can Methods for Comparing Means Have Relatively Low Power, and What Can You Do to Correct the Problem?

Rand R. Wilcox

Certainly, one of the most common goals in applied research is comparing two or more groups in terms of some *measure of location*, that is, a quantity intended to represent the "typical" subject or object under study. Of course, the measure of location routinely used is the population mean, μ. If there is no difference between the distributions associated with two or more groups, standard methods for comparing means appear to provide good control over the probability of a Type I error (i.e., concluding the means are different when in fact they are equal). However, if the groups differ in some way, and in fact you should reject the hypothesis that the groups are the same in terms of some measure of location, then using standard methods for comparing means is one of the worst things you could possibly do. In fact, very slight departures from normality can have serious consequences.

In this article, I review the problem that arises in using conventional statistical methods to compare group means and then discuss some solutions. Standard nonparametric methods do not correct the problem, nor do some of the better known improvements for comparing means. There are, however, new methods that can help applied researchers.

Rand R. Wilcox is Professor of Psychology at the University of Southern California. Address correspondence to Rand R. Wilcox, Department of Psychology, University of Southern California, Los Angeles, CA 90089-1061; e-mail: rwilcox@wilcox.usc.edu.

WHY IS THERE A PROBLEM?

For comparing means, Student's *t* test is the most commonly used method, although some researchers might use Welch's[1] method instead. If you whisper "nonnormality," some researchers might respond by using the Mann-Whitney *U* test, but others might argue that Student's *t* test is robust to nonnormality. For many years within the statistical literature, it has been known that all three of these methods have serious practical problems, particularly in terms of power. Improved methods have now emerged and are ready to be used in applied work.

When choosing a procedure for comparing groups, it helps to keep three common goals in mind:

1. Control the probability of a Type I error when the distributions are identical.
2. Compute accurate confidence intervals for the difference between two measures of location when the distributions differ.
3. Achieve reasonably high power when the two groups differ in terms of some measure of location.

Goal 1 has received the most attention, especially within the social sciences. In this regard, Student's *t* test, and its extension to more than two groups, appears to perform very well.[2]

For Goal 2, Student's *t* test appears to perform reasonably when equal sample sizes are used, but for unequal sample sizes, serious problems arise. In particular, Cressie and Whitford[3] described general circumstances under which, no matter how