CHAPTER

2

Statistical Power in Aphasia Research

Nina F. Dronkers
Bix E. Swain

The calculation of statistical power is an often neglected step in scientific research. This is puzzling, because power is directly related to statistical significance, which is what most researchers seek. In this chapter, we present the basic concepts of statistical power with a discussion of its importance and benefits.

DEFINITION

The most typical way of defining statistical power is in terms of the null hypothesis. Under the null hypothesis of a t-test, for example, the means of the groups we are testing are believed to be equal. In other words, we hypothesize that the groups do not differ significantly. With this in mind, we offer the following working definition of statistical power. *The power of a statistical test is the probability that the test will lead to the rejection of the null hypothesis if the null hypothesis is false.*

Let us pretend that we have a crystal ball that can tell us the true outcome of our statistical test. Let us say that our crystal ball tells us that the null hypothesis of our test is false. Under these conditions, the power of our statistical test is the probability that our test will lead to the rejection of the null hypothesis.

IMPORTANCE

A PRIORI SAMPLE-SIZE ESTIMATION

Statistical power can help the experimenter before a study is begun in the estimation of the sample size necessary to achieve the desired significant difference. To appreciate this, it is important to consider that power is related to two other parameters in addition to sample size: the significance criterion (i.e., the alpha level set at the beginning of the study) and the effect size (i.e., the difference found between the groups related to the variance of these groups). If the values for three of the four parameters are known, then solving for the fourth becomes a simple mathematical task. For example, is we wish to conduct a study in which we have established that a power of .80 and a significance criterion of .05 would be desirable (Cohen, 1977; Kraemer and Thiemann, 1987), and if we have some idea of the effect size, we can then determine how large a sample we would need to be able to reject the null hypothesis when the null hypothesis is false. If we should find that 150 subjects

would be necessary, we might decide that the expense of the study would be prohibitive. Such a discovery would save us the resources necessary to conduct a study that would not yield significant differences with a reasonable number of subjects.

Similarly, a great deal of expense can be spared if we find that we do not have to test as many subjects as we thought we might. We may discover that 10 subjects give us as much power as 20 subjects. By stopping at 10, we save ourselves considerable expense.

POST HOC EVALUATION OF POWER AND SAMPLE SIZE

A second application for power analysis is in determining whether adequate power has been obtained or if additional subjects should be tested. The advantage of post hoc analysis is the knowledge of the actual effect size, which may be difficult to estimate before the data are collected. Table 2-1 illustrates this point.

The data in Table 2-1 are part of a larger study being conducted in our laboratory on the structure of the naming process in aphasic patients (Dronkers, Elman, and Wertz, in preparation). The methodology of this particular study is not important for this discussion. Rather, we are interested in the change in power with the addition of more subjects. The data reflect patients' reaction times to target stimuli preceded either by a semantically related (valid) or a semantically unrelated (invalid) prime stimulus. The difference of interest is between the validly and invalidly primed targets. With seven subjects, we find mean reaction times of 739 ms for target stimuli primed by valid primes and 791 ms for targets that are invalidly primed. The t value achieved with these data is 2.11, with

TABLE 2-1. EXAMPLES OF THE RELATIONSHIP BETWEEN SAMPLE SIZE AND STATISTICAL POWER USING REACTION TIME DATA FROM A STUDY ON SEMANTIC PRIMING IN APHASIC PATIENTS

	Validly primed targets	*Invalidly primed targets*			
Sample size	MEAN SD	MEAN SD	*t*	*p*	*Power*
N = 7	739 156	791 157	2.11	.077	.60
N = 8	725 150	777 150	2.48	.041	.79
N = 20*	722 160	767 155	3.72	.002	.99

*Data on 20 "subjects" was obtained by duplicating existing data from 10 subjects.

an associated p value of .077. Because we had initially established that our significance criterion would be .05, we must determine that we do not have enough evidence to reject the null hypothesis and conclude that no differences are present between the means of these two types of primed targets. A power analysis yields a power of .60, well below the .80 level we had originally set.

Given the difference between the means of the two types of primed targets, it appears worthwhile to continue the study and test additional subjects. The next row in Table 2-1 indicates the means and standard deviations for eight subjects, with a t value of 2.48 and a p value of .04. Considering our significance criterion, we determine that we now have enough evidence to reject the null hypothesis and conclude that there are differences in reaction times between the two types of primed targets. A power analysis yields a power of .79, approximating the .80 we originally desired. Had we decided to discontinue the study after only seven subjects, we would have missed an important difference between the means simply because we lacked the power to bring out the difference. We would have falsely retained the null hypothesis, thereby committing a type II error.

The last row in Table 2-1 reflects data from 10 subjects, whose data have been duplicated to illustrate what can occur if the sample size becomes too large. The t value for these 20 "subjects" is 3.72, with a p value of .002 and an associated power of .99. We have achieved greater statistical power, but we have not changed our decision to reject the null hypothesis. In other words, testing additional subjects may give us greater power, but it does not change our decision rule. The time and effort spent in collecting these data could well have been put toward another study, testing other questions.

One may argue that the larger the sample size, the closer one comes to approximating the tendency of the population. While this is true, the increase in power also may reveal differences that are unimportant, clinically insignificant, or simply account for very little of the variance. This is an important consideration, since resources may be wasted testing for differences that have no impact on the research question.

CONCLUSIONS

Statistical power can be a valuable tool in research, both in planning the investigation and in assessing the completed design. A priori estimations of sample size and post hoc evaluations of power can save the experimenter valuable resources as well as reduce the risk of type II errors. These are desirable goals in any investigation.

REFERENCES

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Dronkers, N. F., Elman, R. J., and Wertz, R. T. The structure of naming ability in aphasia. (In preparation).

Kraemer, H. C., and Thiemann, S. (1987). *How many subjects? Statistical power analysis in research.* Newbury Park, CA: Sage Publications.