Chapter 14
The Power of a Sensitivity Analysis and Its Limit

Abstract In an experiment, power and sample size calculations anticipate the outcome of a statistical test that will be performed when the experimental data are available for analysis. In parallel, in an observational study, the power of a sensitivity analysis anticipates the outcome of a sensitivity analysis that will be performed when the observational data are available for analysis. In both cases, it is imagined that the data will be generated by a particular model or distribution, and the outcome of the test or sensitivity analysis is anticipated for data from that model. Calculations of this sort guide many of the decisions made in designing a randomized clinical trial, and similar calculations may usefully guide the design of an observational study. In experiments, the power in large samples is used to judge the relative efficiency of competing statistical procedures. In parallel, the power in large samples of a sensitivity analysis is used to judge the ability of design features, such as those in Chapter 5, to distinguish treatment effects from bias due to unmeasured covariates. As the sample size increases, the limit of the power of a sensitivity analysis is a step function with a single step down from power 1 to power 0, where the step occurs at a value $\widetilde{\Gamma}$ of Γ called the design sensitivity. The design sensitivity is a basic tool for comparing alternative designs for an observational study.

14.1 The Power of a Test in a Randomized Experiment

What is the power of a test?

The power of a statistical test anticipates the judgment that the test will issue when the test is put to use. Conceptually, the power of a test is the probability that the test will recognize that a false null hypothesis is indeed false and reject it. If the test will reject the null hypothesis when the *P*-value is less than or equal to, say, the conventional 0.05 level, then the power of the test is the probability that the *P*-value will be less than or equal to 0.05 when the null hypothesis is indeed false.

P.R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
DOI 10.1007/978-1-4419-1213-8_14,

If the test is a valid test of its null hypothesis, and if the null hypothesis were true, the probability that the P-value would be less than or equal to 0.05 is itself less than or equal to 0.05. This is the defining property of a valid test or P-value. The P-value is unlikely to be less than or equal to 0.05 if the null hypothesis is true — it happens in only one experiment in 20. Having defined a test so that rejection of a true null hypothesis is improbable, we now want to make rejection of a false null hypothesis highly probable; that is, we would like to have a powerful test.

The power of a test depends upon many things. First and foremost, it depends upon what is true. If the null hypothesis is false, something else is true instead. If the null hypothesis is false but barely so, the power is likely to be little better than chance, 0.05. If the null hypothesis is far from the truth, the power is likely to be much higher. The power depends also upon the sample size, the duration of follow-up, the specifics of the experimental design, and the procedures used in statistical analysis.

Power is a basic tool in designing a randomized experiment. How many patients are needed? Well, calculate the power with 100, 200 and 300 patients. Is it better to study the survival of 200 patients for five years or 300 patients for three years? Calculate the power in each situation. The new treatment is expensive and difficult to apply. To reduce cost, the experimenter is considering randomizing one-third of the patients to treatment and two-thirds to control. Will this design be substantially inferior to randomizing half the patients to treatment and half to control? Calculate the power for both designs. There are two ways to measure the response, one precise but expensive, the other imprecise but inexpensive. If the inexpensive device is used, the money saved will permit a larger sample size. For fixed total cost, which is better, more precision with fewer patients, or less precision with more patients? Calculate the power in both situations. With 20 schools, it would be convenient to randomly assign ten schools to treatment, the rest to control, and to use the same treatment for the hundreds of students in each school. Is that a terrible idea? Would it be vastly better to randomize the many hundreds of students as individuals? Calculate the power for the two designs.

Power is also a basic tool in evaluating the statistical methods that are used to analyze the results of a randomized experiment. Faced with the same data from the same statistical models, two different tests will typically have different power. The power of two different tests in a variety of circumstances is often the basis for choosing which test to use. The t-test has slightly better power than the Wilcoxon test for data from a Normal distribution, but substantially inferior power for distributions with longer tails, which is a basic reason that the Wilcoxon test is preferred; its power is robust.

The remainder of this section briefly reviews the idea of power and its computation for Wilcoxon's signed rank statistic when used in a randomized experiment. The concept of power is important throughout Part III. The details of the computation of power are relevant to the details of the discussion in Part III, but the concepts of Part III should be accessible if the concept of power is clear.