EVAL 6970: Experimental and Quasi-Experimental Designs for Applied Research and Evaluation

Power and Precision

Power Analysis

Traditionally, data collected in a research study is submitted to a significance test to assess the viability of the null hypothesis. The p-value, provided by the significance test and used to reject the null hypothesis, is a function of three factors: size of the observed effect, sample size, and the criterion required for significance (alpha).

A power analysis, executed when the study is being planned, is used to anticipate the likelihood that the study will yield a significant effect and is based on the same factors as the significance test itself. Specifically, the larger the effect size used in the power analysis, the larger the sample size; the larger (more liberal) the criterion required for significance (alpha), the higher the expectation that the study will yield a statistically significant effect.

These three factors, together with power, form a closed system—once any three are established, the fourth is completely determined. The goal of a power analysis is to find an appropriate balance among these factors by taking into account the substantive goals of the study, and the resources available to the researcher.

Effect Size

The term effect size refers to the magnitude of the effect under the alternate hypothesis. The nature of the effect size will vary from one statistical procedure to the next (it could be the difference in cure rates, or a standardized mean difference, or a correlation coefficient), but its function in power analysis is the same in all procedures.

The effect size should represent the smallest effect that would be of clinical or substantive significance, and for this reason, it will vary from one study to the next. In clinical trials, for example, the selection of an effect size might take into account the severity of the illness being treated (a treatment effect that reduces mortality by 1% might be clinically important, while a treatment effect that reduces transient asthma by 20% may be of little interest). It might take into account the existence of alternate treatments. (If alternate treatments exist, a new treatment would need to surpass these other treatments to be important.) It might also take into account the treatment's cost and side effects. (A

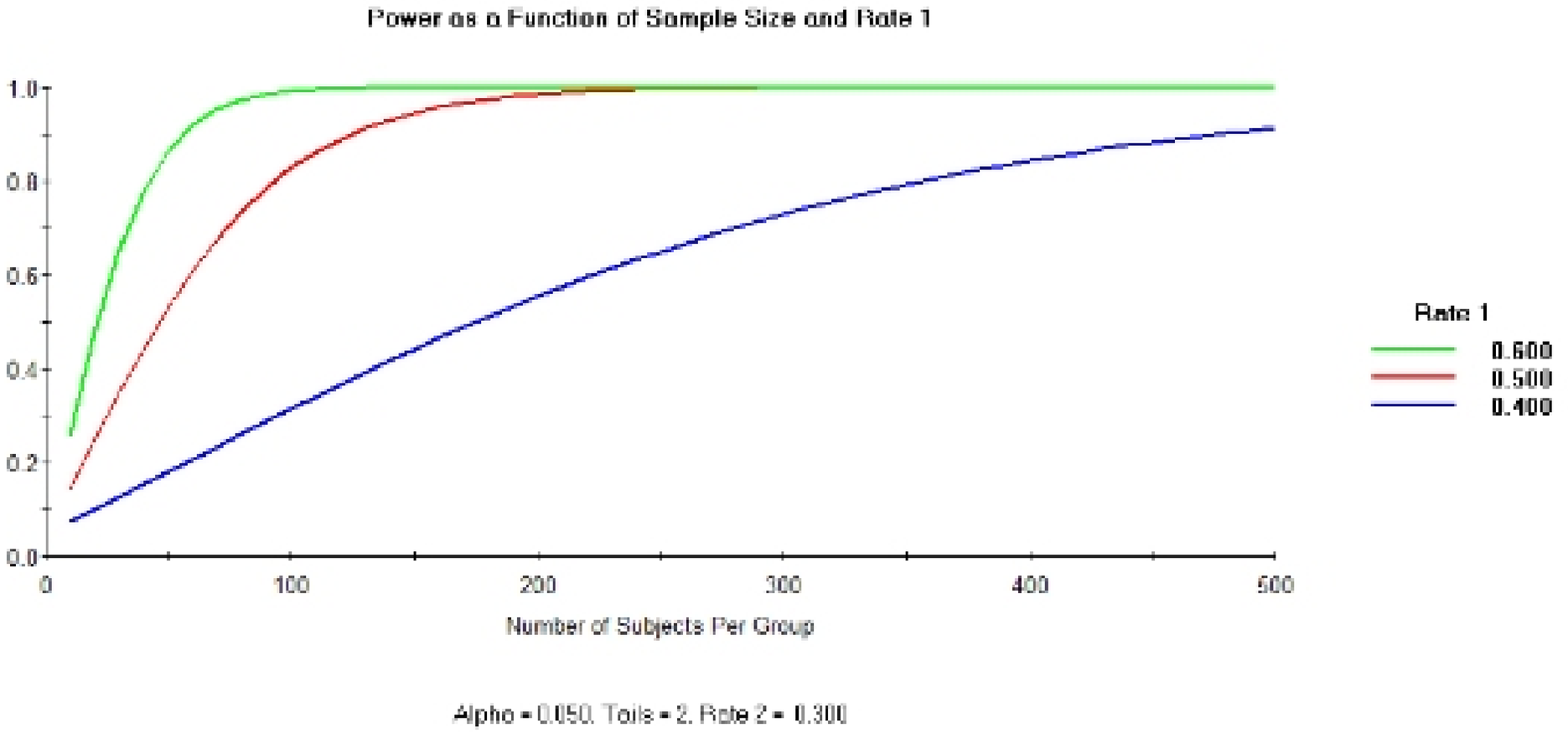
treatment that carried these burdens would be adopted only if the treatment effect was very substantial.)

Power analysis gives power for a specific effect size. For example, the researcher might report that if the treatment increases the recovery rate by 15 percentage points, the study will have power of 80% to yield a significant effect. For the same sample size and alpha, if the treatment effect is less than 15 percentage points, then the power will be less than 80%. If the true effect size exceeds 15 percentage points, then power will exceed 80%.

While one might be tempted to set the "clinically significant effect" at a small value to ensure high power for even a small effect, this determination cannot be made in isolation. The selection of an effect size reflects the need for balance between the size of the effect that we can detect and resources available for the study.

Small effects will require a larger investment of resources than large effects. Figure 1 shows power as a function of sample size for three levels of effect size (assuming that alpha, two-tailed, is set at 0.05). For the smallest effect (30% versus 40%), we would need a sample of 356 per group to yield power of 80%. For the intermediate effect (30% versus 50%), we would need a sample of 93 per group to yield this level of power. For the largest effect size (30% versus 60%), we would need a sample of 42 per group to yield power of 80%. We may decide that for our purposes, it would make sense to enroll 93 per group to detect the intermediate effect but inappropriate to enroll 356 patients per group to detect the smallest effect.

Figure 1



The true (population) effect size is not known. While the effect size used for the power analysis is assumed to reflect the population effect size, the power analysis is more appropriately expressed as, "If the true effect is this large, power would be …," rather than, "The true effect is this large, and therefore power is …."

This distinction is an important one. Researchers sometimes assume that a power analysis cannot be performed in the absence of pilot data. In fact, it is usually possible to perform a power analysis based entirely on a logical assessment of what constitutes a clinically (or theoretically) important effect. Indeed, while the effect observed in prior studies might help to provide an estimate of the true effect, it is not likely to be the true effect in the population—if we knew that the effect size in these studies was accurate, there would be no need to run the new study.

Since the effect size used in power analysis is not the true population value, the researcher may decide to present a range of power estimates. For example (assuming that $N = 93$ per group and alpha = .05, two-tailed), the researcher may state that the study will have power of 80% to detect a treatment effect of 20 points (30% versus 50%) and power of 99% to detect a treatment effect of 30 points (30% versus 60%).

Cohen has suggested conventional values for small, medium, and large effects in the social sciences. The researcher may want to use these values as a kind of reality check to ensure that the values that he or she has specified make sense relative to these anchors.

Alpha

The significance test yields a computed *p*-value that gives the likelihood of the study effect, given that the null hypothesis is true. For example, a *p*-value of 0.02 means that, assuming that the treatment has a null effect, and given the sample size, an effect as large as the observed effect would be seen in only 2% of studies.

The *p*-value obtained in the study is evaluated against the criterion, alpha. If alpha is set at 0.05, then a *p*-value of 0.05 or less is required to reject the null hypothesis and establish statistical significance.

If a treatment really is effective and the study succeeds in rejecting the nil hypothesis, or if a treatment really has no effect and the study fails to reject the nil hypothesis, the study's result is correct. A Type I error is said to occur if there is a nil effect but we mistakenly reject the null. A type 2 error is said to occur if the treatment is effective but we fail to reject the nil hypothesis.

Note: The null hypothesis is the hypothesis to be nullified. When the null hypothesis posits a nil effect (for example, a mean difference of 0), the term nil hypothesis is used. Assuming that the null hypothesis is true and alpha is set at 0.05, we would expect a Type I error to occur in 5% of all studies—the Type I error rate is equal to alpha. Assuming that the null hypothesis is false (and the true effect is given by the