Power calculations D.A. Freedman December 2007

What is the power of a randomized controlled experiment? We assume there are $n_1 + n_2$ subjects, of whom $n_1$ are assigned at random to treatment, the rest being controls. Typically, $n_1$ and $n_2$ are determined by cost considerations. We take these as given.

We model treatment data as observed values of $n_1$ IID random variables $Y_1, \ldots, Y_{n_1}$. We model the control data as observed values of $n_2$ IID random variables $X_1, \ldots, X_{n_2}$. Thus, we are pretending to have random samples from two infinite populations. We assume common variance $\sigma^2$, different population means $\mu_1$ and $\mu_2$.

The null hypothesis is $\mu_1 = \mu_2$. Assume arguendo that we make a two-sided 5% test of the null against the alternative $\mu_1 \neq \mu_2$, and we compute power at $\mu_1 = \mu_2 + k\sigma$ where $k > 0$ is chosen so that an effect of $k\sigma$ is of substantive interest.

We assume $n_1$ and $n_2$ are so large that the CLT applies, and $\hat{\sigma} \doteq \sigma$, where $\hat{\sigma}^2$ is the usual pooled estimator of variance. Let

$$f = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

The test statistic is

$$z = (\overline{Y} - \overline{X})/s, \text{ where } s = \hat{\sigma} f \doteq \sigma f$$

We reject when $|z| > 2$. The size is about 5%, as required. Under the alternative, $z$ is asymptotically distributed as $N(k^*, 1)$, where

$$k^* = k/f$$

Suppose $k^* > 2$. If $Z$ is $N(0, 1)$, asymptotic power is then

$$P(Z > 2 - k^*) + P(Z < -2 - k^*),$$

the second term being negligible. Small changes in $k^*$ can make big changes in apparent power.

It is advantageous to get power around 80%, which corresponds to $k^* \doteq 2.85$. Set $k^* = 2.85$ and then $k = k^* f$. If you don't like the answer, switch to one-sided null hypothesis and one-sided test; the multiplier changes from 2 to 1.65, so power is essentially $P(Z > 1.65 - k^*)$, and $k^* = 2.5$ is a good $k^*$.

If power is still insufficient, you have to increase the sample sizes. This is the real way to increase power, but it's slow going until you get near the sweet spot: doubling sample sizes only reduces the SE for the sample difference by a factor of $\sqrt{2}$.

If power is around 50%, granting agencies may view the experiment as a crapshoot, unworthy of support. If power is over 95%, the experiment may be viewed as sure to give a positive result, hence a boondoggle, hence unworthy of support. These are mistakes, but natural ones. Different fields of application are likely to have different conventions about the "right" power in a grant application. I think that in medical research, 80%–90% is the right range,

> Rule of thumb. Guesstimate the SE for the difference between the sample means. If you assume the difference between the population means is 2 SE, you will have minimal power. At 3 SE, you will have oodles of power.

If sample sizes are below 15, say, don't use the CLT. Assume the parent populations are normal and use the $t$ distribution rather than the normal to get the cutoff (like 2) for 5%. There will be $n_1 + n_2 - 2$ degrees of freedom.

Similar calculations can be done for more complex designs, but, they get more complex. Similar techniques also work for statistics like (log) odds ratios. To get started, see

http://www.stat.berkeley.edu/users/census/oddsrat.pdf

Blocking and stratification will reduce variance, i.e., increase power. Heterogeneity (subject-to-subject variation in treatment effects) will increase variance, i.e., reduce power. Using $t$ will reduce power, although not much if $n_1 + n_2 > 20$, say. One-sided tests give more (apparent) power.

The modeling assumptions are a little goofy, but, all you're doing is ticking a box.